The international journal of science / 25 June 2020

nature

GAIN OF FUNCTION

Astrocytes rewired to replace neurons lost in Parkinson's disease

Site loss
Israel's archaeological heritage is vanishing under concrete

Modellers' manifesto
Show your workings, admit your politics and share your blindspots

Security key
Satellites enable long-range quantum cryptography

Vol. 582, No. 7813
nature.com

The international journal of science / 25 June 2020

nature

Leaving the WHO might not be as easy as Trump thinks

There is no process for member countries to withdraw. The United States should stay and help to reform the agency from within.

US President Donald Trump's decision, announced on 29 May, to withdraw funding from the World Health Organization (WHO) was never in doubt.

Since the beginning of the coronavirus outbreak, the White House has been intensifying its charge that the WHO was slow to respond to the threat, and overly influenced by China. Undoubtedly, the agency has lessons to learn, and, at the World Health Assembly last month, WHO member states endorsed an independent evaluation. It is irresponsible and dangerous for the United States — the WHO's largest donor — to bypass the agreed process and withhold roughly US$450 million in annual funding in the middle of one of the worst pandemics in recent history. This will undermine the world's efforts to control the new coronavirus and will endanger more lives as COVID-19 continues on its destructive path.

The chances that the US decision could be reversed at this stage are slim, but there is a small window of time, during which everything possible must be tried. There is too much at stake not to do so.

The United States was among the biggest champions for establishing an international agency to assist countries in rebuilding national health systems after the Second World War. Since then, the United States' compulsory and voluntary financial contributions have enabled the WHO to carry out life-saving work in low- and middle-income countries and regions — for example, in treating Ebola, HIV and polio. And US public-health researchers and policymakers are embedded in the organization's many research and policy-making bodies.

US researchers are also members of many of the WHO's scientific-advisory groups, including those on COVID-19. And US institutions, especially the Centers for Disease Control and Prevention, work with the WHO by hosting what are called collaborating centres. One such centre, which has partners in Australia, China, Japan and the United Kingdom, monitors influenza and helps to design flu vaccines.

The US presence in the WHO was important to the reform, in 2005, of the International Health Regulations, under which countries are obliged to accurately report outbreaks, cases and deaths. The regulations needed strengthening because under-reporting had been a feature of past disease outbreaks, resulting in lost lives. The authors of these regulations could not have imagined that a US president would promote and then justify collecting inaccurate data. Trump did so at a rally on 20 June, when he said he had asked officials to go slow on coronavirus testing — a clear breach of these regulations.

> "The WHO's founding constitution lacks a provision for countries to withdraw."

Researchers are debating what form the US withdrawal will take, and how quickly it could happen. Funding that has been given cannot be taken back; nor can voluntary contributions that have been pledged in advance. And the WHO's founding constitution lacks a provision for countries to withdraw. Under a resolution passed by both US houses of Congress in 1948, the United States must give one year's notice and pay any outstanding funds if it wishes to leave. Whether the White House will be bound by this, and what powers Congress has to enforce its earlier decision, are a matter of debate. But, as far as the WHO constitution is concerned, countries that join remain members.

Those interviewed for this editorial — researchers in international law and public health, WHO advisers and members of other multilateral processes — agree that a member state cannot be compelled to stay. The Soviet Union famously led a walkout of Eastern bloc countries from the WHO in 1949 owing to concerns that the United States was too dominant, and these countries returned only after the death of Joseph Stalin in 1953. But the absence of a formal withdrawal mechanism allowed the WHO's first director-general, Canada's Brock Chisholm, to classify the Soviet membership as 'inactive' rather than 'withdrawn'. Something similar could happen now, creating a path for the eventual return of the United States should it leave.

The United Nations did eventually create rules — the 1969 Vienna Convention on the Law of Treaties and its 1986 extension — that cover how countries can exit a multilateral organization from which there is no provision to withdraw. But the United States is among those that have not ratified these agreements, a decision that the White House might come to regret.

Leave to remain

Trump's decision to withdraw from the WHO seems to follow a pattern of behaviour that includes the 2017 decision to leave the UN's science-cooperation agency UNESCO, and his ending of US involvement in the Iran nuclear deal. But this latest move is different in one important respect: some influential voices in Trump's own Republican Party are urging him to reconsider. That represents an audience for researchers, research institutions, industry and health campaigners to work with, to highlight the dangers of a US exit. Lawmakers must be pressed to reverse this dangerous decision, or, at the very least, to ensure that any outstanding dues are paid and that the one-year period of notice before withdrawal is respected.

If the United States wants to improve the WHO, it needs to back both the independent evaluation, as other WHO member states have done, and implementation of recommended changes — not turn its back entirely. If, as seems probable, the Trump administration does order a swift withdrawal, the WHO's constitutional duty is to keep the country's seat, so that the United States can quickly return when a future leader makes a wiser choice.

nature

When the White House knew how to do diplomacy

The race to sequence the human genome ended in a tie 20 years ago, thanks to some deft statecraft by the Clinton administration.

On 26 June 2000, US President Bill Clinton and UK Prime Minister Tony Blair presided over a carefully choreographed piece of scientific theatre. Through a video link connecting Washington DC and London, they announced to the world that scientists had completed a rough first draft of the human genome sequence.

It was quite a production. Amid accompanying music and applause from scientists, diplomats and members of Clinton's cabinet, the president entered the White House East Room. He was flanked by the two leaders of competing teams on the sequencing effort: Francis Collins, then-director of the US National Human Genome Research Institute, and Craig Venter, founder of Celera Genomics, a company formed to commercialize genome data.

It was not a day for understatements, as the reporter covering the event for *Nature* wrote. One participant, Mike Dexter, then-director of the Wellcome Trust, described its significance as surpassing that of the invention of the wheel. Clinton himself said: "Today's announcement represents more than just an epoch-making triumph of science and reason … With this profound new knowledge, humankind is on the verge of gaining immense, new power to heal."

Exactly 20 years on from that event, the ground-breaking significance of determining the human genome sequence is clear: it sparked a revolution in human biology and medicine, and genome sequencing is now routine.

No winners or losers

Less has been said about how the start of biology's new era marked the culmination of one of the last great contests of twentieth-century science. It is hard to imagine today's politicians and their advisers declaring a truce between duelling scientists — or reminding scientists that cooperation has as much value as competition. Clinton was keen to stress that there would be no winners or losers from the sequencing race. "From this moment forward, the robust and healthy competition that has led us to this day … will be coupled with enhanced public–private cooperation," he said, after which all three men — Clinton, Collins and Venter — shook hands.

The roots of the two teams' rivalry can be traced back to the early 1990s, when Venter resigned from his post as a researcher at the US National Institutes of Health (NIH) in Bethesda, Maryland, to work full time on establishing genome-sequence-data businesses. Earlier, in 1990, researchers and public funding agencies in the United States had launched the Human Genome Project (HGP), an international consortium committed not only to genome sequencing, but also to ensuring that its data would be free for researchers to access.

> "It is hard to imagine today's politicians reminding scientists that cooperation has as much value as competition."

At a meeting in Bermuda in February 1996, the HGP's partners agreed to release sequence data every 24 hours and to deposit these data in public databases. Venter declined to be a part of this arrangement and the two groups found themselves in open dispute. Venter argued that the HGP was spending scarce public funds — some US$3 billion — on a cumbersome approach to sequencing requiring "armies of scientists" with little scope for innovation. Meanwhile, members of the HGP questioned the ethics of Venter's business model.

Peace talks

Attempts were made to broker peace and foster cooperation, but they ended in failure and acrimony. As late as March 2000, when talks between the two sides broke down, Venter told reporters that the HGP's decision to release the text of a letter it had sent to Celera outlining what it saw as sticking points was "a low-life thing to do". One leading member of the HGP, John Sulston, then-director of the Sanger Centre (now the Wellcome Sanger Institute) in Hinxton, UK, said Celera's taking of public data and selling it along with their own amounted to a "con-job".

The extent of the vitriol on a flagship US science project did not go down well with the White House, and Neal Lane, Clinton's chief science adviser, who is now at Rice University in Houston, Texas, says that the president pressed for the dispute to be resolved. But, all the while, both sides — including the more than 1,000 researchers involved in the public effort — were continuing with their sequencing work. With a completed sequence in sight, the two groups eventually agreed that they would cross the finishing line together — and Celera would publish its sequence in the scientific literature. In his White House statement, Clinton repaid the gesture by declaring support for biotechnology companies and for the patenting of genetic discoveries.

The eventual agreement was brokered principally by Ari Patrinos at the US Department of Energy — where the idea to sequence the genome had originated in the 1980s — and Eric Lander at the Whitehead Institute at the Massachusetts Institute of Technology in Cambridge, which hosted one of the HGP sequencing centres. Patrinos invited Venter and Collins to meet at his house over pizza. "It was just the three of us; it was amazing how quickly the ice melted," Patrinos later said.

Looking back at the 40-minute announcement, the fact that world leaders played a part in efforts to tie the race to sequence the human genome is striking. It also serves as an unhappy reminder that, although biology has continued to progress, standards of statesmanship have fallen to previously unimaginable depths.

It is hard to imagine Donald Trump or Boris Johnson having such a role today.

World view

By Ann Pettifor

Rebuild the ramshackle global financial system

Economic researchers neglect the role of financialization in global existential crises.

Riddled with comorbidities, the current global monetary and financial set-up precipitates crises with increasing frequency. At first, these were on the fringes of the global economy; in 2007–09 they moved to its very core.

Since 1971, national economies, and all our lives, have been shaped by this 'system', which can be described only as ramshackle. In that year, US president Richard Nixon unilaterally dismantled the Bretton Woods international financial architecture, built at the end of the Second World War. No sound replacement was constructed. Largely privatized, what governs the global economy today is mostly deregulated and made up of an ad hoc set of legal arrangements.

The financial system's role in driving global connectedness has led to many changes: radical innovations in information and transport technologies; the greater integration of trade and business; and rises in living standards. However, the system has resulted in greater fragility. Financialization of the global economy plays a part in transmitting pathogens by financing long supply lines and international transport networks. These impacts are barely understood by policymakers, and are rarely discussed by mainstream economists.

Instead, academics are preoccupied with theories centred on the nation-state. Most focus primarily on microeconomics — the study of individuals, households and firms — and their relation to government. Too few tackle macroeconomic questions of cross-border capital flows; the role of central bankers in overseeing and making global financial markets; or the dominance of the US Federal Reserve in global governance.

Nor is there sufficient interest in the US dollar's often maligned role as the world's reserve currency, which must be held by many countries to make payments to other countries. When speculators 'fly' from investment in one country and convert their holdings into US dollars, the original currency plummets, raising the cost of imports such as oil or pharmaceuticals. This is not new. Countries are periodically rocked by stampedes of such cross-border capital flows. The 'exorbitant privilege' exercised by the US dollar accelerates such stampedes. Action by the Federal Reserve can only briefly ameliorate these shocks, because of this volatility.

New lines of research into financial globalization are needed to manage domestic economies in these challenging times and mitigate the impact of future crises — from pandemics to climate breakdown and biodiversity collapse. Among the most important studies will be those leading to the development of a new, better-managed international financial architecture.

> **"The stimuli have once again bailed out monetary chicanery more than they have individuals."**

Ann Pettifor is an economist and government adviser in London. Her latest book is *The Case for the Green New Deal*. e-mail: ann.pettifor@primeeconomics.org

The possibility of such a transformation to ensure stability and sustainability is largely absent from academic and public discourse. The 'rethinking' after the 2007–09 global financial crisis led simply to a consolidation of the existing system. As the International Monetary Fund (IMF) explained in its April 2020 Global Financial Stability Report, after that crisis, the US Federal Reserve, far from limiting excessive credit creation, turned a blind eye as private credit markets expanded and reached US$9 trillion globally. As with the 2007–09 crisis, lax regulation by central bankers lowered borrowers' credit quality, and weakened underwriting standards and investor protections. These risky credit markets — in high-yield ('junk') bonds, leveraged loans and private debt — showed stresses through early April 2020, despite the Federal Reserve's massive cash injection. The stimuli have once again bailed out monetary chicanery more than they have individuals.

Reform is crucial if societies north and south are to mobilize the financing needed to import drugs and equipment to tackle this pandemic and future ones, let alone if they are to fight climate change by investing public resources into alternative and sustainable transport, land use and energy systems. During the COVID-19 crisis, private markets failed to supply timely necessities, such as personal protective equipment. In the same way, before this pandemic, the market proved unable to provide affordable health care, housing and higher education, as well as decent well-paid jobs, for millions of citizens. Private markets are not fit to guarantee the security of populations in the face of increasingly dangerous extreme weather events.

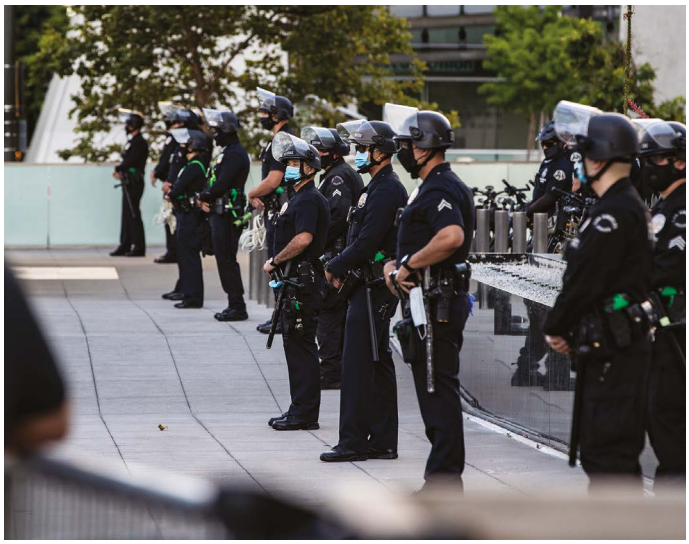
Just as we need a sustainable global ecosystem, so do we need a stable, sustainable international financial system — an important public good. The next practical priority, therefore, is to ensure international coordination and collaboration in designing a new architecture.

When the international system was last reconstructed at Bretton Woods, New Hampshire, in 1944, US president Franklin D. Roosevelt invited only qualified economists to the conference. Bankers and financiers were barred. And the selection of experts was broad — not narrowly focused on Anglo-American academia. Roosevelt ensured that scholars represented diverse geographical regions and interests: 32 of the 44 delegations came from low-income countries. As political scientist Eric Helleiner explained in his 2014 book *Forgotten Foundations of Bretton Woods*, policymakers were "deeply committed to an inclusive 'procedural multilateralism'". It gave a formal voice to all the United Nations, and to other nations that had remained neutral.

Now, as then, we need international, pluralistic academic and political leadership to foster a new procedural multilateralism. That is the only way to build a global monetary system that can help countries to end this pandemic — and to tackle climate breakdown.

BOB FOYERS PHOTOGRAPHY

News in brief



MATHEMATICIANS URGE COLLEAGUES TO BOYCOTT POLICE WORK IN WAKE OF KILLINGS

A group of mathematicians in the United States has written a letter calling for their colleagues to stop collaborating with police because of the widely documented disparities in how US law-enforcement agencies treat people of different races and ethnicities. They concentrate their criticism on predictive policing, a maths-based technique aimed at stopping crime before it occurs.

The letter, dated 15 June, is addressed to *Notices of the American Mathematical Society* (AMS), and comes in the wake of protests in the United States and globally, sparked by the killing of George Floyd by a police officer in Minneapolis, Minnesota, in May. More than 1,400 researchers have signed it (see go.nature.com/3fkwxmb).

In recent years, mathematicians, statisticians and computer scientists have been developing algorithms that crunch data and claim to help police reduce crime — for instance, by suggesting where crime is most likely to occur and focusing more resources in those areas. Software based on such algorithms is used by police departments across the United States, but many contest its effectiveness.

"Given the structural racism and brutality in US policing, we do not believe that mathematicians should be collaborating with police departments in this manner," the mathematicians write. "It is simply too easy to create a 'scientific' veneer for racism."

"The activity of collaborating with the police is not something we feel a mathematician should be doing," says co-author Jayadev Athreya, a mathematician at the University of Washington in Seattle. (He and the other writers emphasize that the letter represents their own views and not those of their employers.)

The AMS says that it "has no official position on mathematicians' involvement in providing expertise to law-enforcement agencies, or to companies that do business with such agencies".

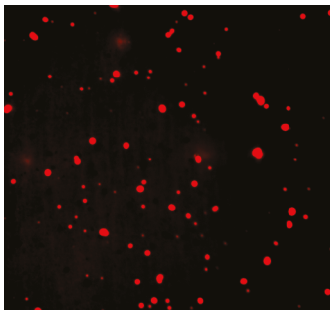
L TO R: GINA FERAZZI/*LOS ANGELES TIMES* VIA GETTY; POLAR MEDIA; ISAAC KLEIN/WHITEHEAD INSTITUTE

CERN PUSHES TO BUILD €21-BILLION SUPERCOLLIDER

CERN — Europe's pre-eminent particle-physics organization — has taken a big step towards building a 100-kilometre circular supercollider to push the frontier of high-energy physics. The decision to pursue the machine was unanimously endorsed by the CERN Council, the organization's governing body, on 19 June, following the plan's approval by an independent panel in March.

The accelerator is expected to cost at least €21 billion (US$24 billion) and would be a follow-up to the Large Hadron Collider. It would be built in an underground tunnel near CERN's location outside Geneva, Switzerland, and smash together electrons and their antimatter partners, positrons, by the middle of the century. The design will enable physicists to study the properties of the Higgs boson and, later, to host an even more-powerful machine that will collide protons in the second half of the century.

The decision is not yet a final go-ahead, and CERN will need global help to fund the project. But the organization can now focus on designing the collider and researching its feasibility. "I think it's a historic day for CERN and particle physics, in Europe and beyond," said CERN director-general Fabiola Gianotti.



'LAVA LAMP' EFFECT COULD MAKE DRUGS MORE POWERFUL

Researchers have used a phenomenon called phase separation to concentrate cancer-drug compounds into precise spots within cells — a discovery that could boost drug-development efforts.

Like blobs in a lava lamp or oil shaken in water, cell components such as proteins and RNA can self-organize into liquid-like droplets known as condensates. Research by a team at the Whitehead Institute in Cambridge, Massachusetts, reveals that synthetic compounds can be sequestered in droplets in a similar way. The effect could be exploited to make certain drugs hit their targets more effectively, while limiting harmful side effects.

The researchers tracked the dynamics of five cancer drugs in test-tube experiments and in human cancer cells in culture (I. A. Klein *et al., Science* http://doi.org/dz3g; 2020). The group found that when mixed with proteins known to form condensates, drugs such as cisplatin cluster into highly concentrated droplets (pictured).

The work could also help efforts to find drugs to fight COVID-19. In unpublished work, the same team has found that proteins from the coronavirus clump together in condensates that can absorb and concentrate drug compounds.

News in focus

MICHAEL DANTAS/AFP/GETTY

Many countries are struggling to count all their coronavirus deaths.

HOW DEADLY IS THE CORONAVIRUS? SCIENTISTS ARE CLOSE TO AN ANSWER

Researchers use the infection fatality rate to gauge how to respond to a new disease, but it's tricky to calculate during an outbreak.

By Smriti Mallapaty

One of the most crucial questions about an emerging infectious disease such as the new coronavirus is how deadly it is. After months of collecting data, scientists are closing in on an answer — in the form of a figure known as the infection fatality rate (IFR). This metric gives the proportion of infected people who will die as a result of a disease, including those who don't get tested or show symptoms.

"The IFR is one of the important numbers alongside the herd-immunity threshold, and has implications for the scale of an epidemic and how seriously we should take a new disease," says Robert Verity, an epidemiologist at Imperial College London.

Calculating an accurate IFR is challenging in the middle of any outbreak because it relies on knowing the total number of people infected — not just those who are confirmed through testing. But the fatality rate is especially difficult to pin down for COVID-19, says Timothy Russell, a mathematical epidemiologist at the London School of Hygiene and Tropical Medicine. That's partly because there are many people with mild or no symptoms, whose infection has gone undetected, and also because the time between infection and death can be as long as two months. Many countries are also struggling to count all their virus-related deaths, he says. Death records suggest that some are being missed in official counts.

Data from early in the pandemic

overestimated the deadliness of SARS-CoV-2, the virus that causes COVID-19, and then later analyses underestimated its lethality. Now, numerous studies — using a range of methods — estimate that in many countries, 5–10 people will die for every 1,000 people with COVID-19. "The studies I have any faith in are tending to converge around 0.5–1%," says Russell.

But some researchers say that convergence between studies could just be coincidence. For a true understanding of how deadly the virus is, scientists need to know how readily it kills different groups of people. The risk of dying from COVID-19 can vary considerably, depending on age, ethnicity, access to health care, socio-economic status and underlying health conditions. More high-quality surveys of different groups are needed, these researchers say.

IFR is also specific to a population and changes over time as doctors get better at treating the disease, which can further complicate efforts to pin it down.

Getting the number right is important because it helps governments and individuals to determine appropriate responses. "Calculate too low an IFR, and a community could under-react, and be underprepared. Too high, and the overreaction could be at best expensive, and at worst [could] also add harms from the overuse of interventions like lockdowns," says Hilda Bastian, who studies evidence-based medicine at Bond University in the Gold Coast, Australia.

Bridging the gap

Some of the first indications of the virus's deadliness were gleaned from the total number of confirmed cases in China. In late February, the World Health Organization crudely estimated that 38 people had died for every 1,000 with confirmed COVID-19 diagnoses. The death rate among these people — known as the case fatality rate — reached as high as 58 out of 1,000 in Wuhan, the city where the virus emerged. But such estimates exaggerated the disease's deadliness because they did not account for the many people who had the virus but were not tested, obscuring the outbreak's true spread.

Researchers tried to address this gap by estimating the IFR from models that projected the virus's spread. The result from these early analyses hovered around 0.9% — 9 deaths for every 1,000 people infected — with a broader range of 0.4–3.6%, says Verity. His own modelling estimated an overall IFR for China of 7 deaths for every 1,000 people infected, increasing to 33 per 1,000 among those aged 60 or older[1] (see 'How deadly is SARS-CoV-2?').

Russell's team also used data gathered from a large COVID-19 outbreak on the *Diamond Princess* cruise ship in early February to estimate an IFR in China. Almost all of the 3,711 passengers and crew members were tested, enabling researchers to count the total number of infections, including asymptomatic ones, and deaths in a known population. From this, the team estimated an IFR of 0.6%, or 6 deaths for every 1,000 infected people[2].

"The intention of these studies was to gain some ball-park estimates of how deadly COVID-19 is," says Verity.

But researchers also had to make complicated estimates, which still need to be verified, about the number of confirmed cases and the actual number of infected people. "There is value to obtaining rapid early estimates of the IFR, [but] these should be updated as a matter of urgency once better data becomes available," says Verity.

Widespread population surveys that test people for antibodies to the virus, known as seroprevalence surveys, were expected to help refine IFR estimates even further. About 120 such surveys are under way worldwide.

"There is value to obtaining rapid early estimates of the infection fatality rate."

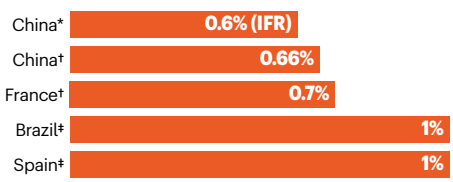
But results from the first antibody studies only muddied the water, suggesting that the virus was less deadly than previously thought. "It got a bit messy," says Russell.

One of the earliest studies tested 919 people in the German town of Gangelt, where a large outbreak had occurred[3]. Of these people, about 15.5% had antibodies against the virus — 5 times higher than the percentage of people known to have had COVID-19 in the town at the time. The figure was used to estimate an IFR of 0.28%. But researchers noted that the study was based on a relatively small number of people.

Other early seroprevalence studies did not properly account for the lack of sensitivity and specificity in the antibody test kits that were used, or for discrepancies between the sampled and underlying populations, says Verity.

HOW DEADLY IS SARS-COV-2?

The infection fatality rate (IFR) is the proportion of people with COVID-19 who will die from the disease. Estimates are for specific regions, and can vary depending on demographics, health-care access and study methodology.



*Estimate based on natural experiment. †Estimate based on modelling. ‡Estimate based on prevalence data.

These issues could have inflated estimates of the total number of infected people and so made the virus seem less deadly, he says. Equally, if COVID-19 deaths go undetected — a problem in many countries that aren't testing all deceased people for the virus — that, too, can bias the fatality rate, says Gideon Meyerowitz-Katz, an epidemiologist at the University of Wollongong, Australia.

Some larger seroprevalence studies have emerged in recent weeks, and these estimate a higher fatality rate than do early studies. One survey[4], posted on medRxiv, of more than 25,000 people across Brazil, estimated an IFR of 1%.

Another survey that tested more than 60,000 people across Spain reports a prevalence of 5%, although the results have not been formally analysed (see go.nature.com/2brqo2c). The survey team did not calculate a fatality rate itself but, on the basis of the results, Verity estimates that Spain has an IFR of around 1% — or 10 deaths for every 1,000 infected individuals.

Several researchers, including Russell and Verity, find it interesting that a growing number of studies from different regions have estimated IFRs in the range of 0.5–1%. But other scientists are cautious about suggestions of agreement. "The trend is potentially more luck than anything else," says Meyerowitz-Katz.

Marm Kilpatrick, an infectious-disease researcher at the University of California, Santa Cruz, also notes that most of the serological data haven't been published in scientific manuscripts. It's hard to know when and how they were collected, and to properly calculate an IFR that accounts for the delay between people getting infected and dying, he says.

Kilpatrick and others are eagerly awaiting large studies that estimate fatality rates across age groups and among those with pre-existing health conditions, which will provide the most accurate picture of how deadly the disease is. One of the first studies to account for the effect of age was posted on a preprint server on 12 June[5]. The study, based on seroprevalence data from Geneva, Switzerland, estimates an IFR of 0.6% for the total population, and 5.6% for people aged 65 and older.

The results have not been peer reviewed, but Kilpatrick says the study addresses many of the issues in previous seroprevalence surveys. "This study is fantastic. It's precisely what should be done with all of the serological data," he says.

1. Verity, R. *et al. Lancet* **20**, 669–677 (2020).
2. Russell, T. W. *et al. Euro Surveill.* **25**, 2000256 (2020).
3. Streeck, H. *et al.* Preprint at medRxiv https://doi.org/10.1101/2020.05.04.20090076 (2020).
4. Hallal, P. C. *et al.* Preprint at medRxiv https://doi.org/10.1101/2020.05.30.20117531 (2020).
5. Perez-Saez, F. *et al.* Preprint at OSF https://doi.org/10.31219/osf.io/wdbpe (2020).

SOURCES: CHINA*: REF. 2; CHINA†: REF. 1; FRANCE: H. SALJE *ET AL. SCIENCE* HTTP://DOI.ORG/DVT3 (2020); BRAZIL: REF. 4; SPAIN: SPANISH MINISTRY OF HEALTH, CONSUMER AFFAIRS AND SOCIAL WELFARE 2020 REPORT

SIMON DAWSON/BLOOMBERG/GETTY

The steroid dexamethasone improves survival in severe cases of COVID-19.

STEROID IS FIRST DRUG SHOWN TO PREVENT DEATHS FROM COVID-19

In a large trial, dexamethasone cut deaths by one-third among critically ill patients.

By Heidi Ledford

An inexpensive and commonly used steroid can save the lives of people seriously ill with COVID-19, a randomized, controlled clinical trial in the United Kingdom has found. The drug, called dexamethasone, is the first shown to reduce deaths from the coronavirus that has killed more than 440,000 people globally. In the trial, it cut deaths by about one-third in patients who were on ventilators because of coronavirus infection.

"It's a startling result," says Kenneth Baillie, an intensive-care physician at the University of Edinburgh, UK, who serves on the steering committee of the trial, called RECOVERY. "It will clearly have a massive global impact." RECOVERY researchers announced the findings in a press release on 16 June, and posted their results in a preprint on 22 June (P. Horby *et al*. Preprint at medRxiv http://doi.org/dz5x; 2020).

RECOVERY, launched in March, is one of the world's largest randomized, controlled trials for coronavirus treatments. The dexamethasone arm enrolled 2,100 participants who received the drug at a low-to-moderate dose of 6 milligrams per day for 10 days, and compared how they fared against about 4,300 people who received standard care for COVID-19.

Dexamethasone's effect was most striking among patients on ventilators. Those who were receiving oxygen therapy but were not on ventilators also saw improvement: their risk of dying was reduced by 20%. The steroid had no effect on people with less severe cases of COVID-19 — those not receiving oxygen or ventilation.

Shortly after the results were released, the UK government immediately authorized the use of dexamethasone for patients hospitalized with COVID-19 who required oxygen, including those on ventilators.

"Finding effective treatments like this will transform the impact of the COVID-19 pandemic on lives and economies across the world," said Nick Cammack, head of the COVID-19 Therapeutics Accelerator at Wellcome, a UK biomedical research charity in London, in a statement. "While this study suggests dexamethasone only benefits severe cases, countless lives will be saved globally."

Rigorous study

Use of steroids to treat viral respiratory infections such as COVID-19 has been controversial, notes Peter Horby, an infectious-disease specialist at the University of Oxford, UK, and a chief investigator on the trial. Data from steroid trials during outbreaks of severe acute respiratory syndrome and Middle East respiratory syndrome caused by related coronaviruses were inconclusive, he says. Nevertheless, given dexamethasone's broad availability, and some promising results from steroid studies in previous outbreaks, RECOVERY investigators considered it important to test the treatment in a rigorous trial, says Horby.

Treatment guidelines from the World Health Organization and many countries have cautioned against treating people with coronavirus with steroids, and some investigators were concerned about anecdotal reports of widespread steroid treatment. The drugs suppress the immune system, which could provide some relief for patients whose lungs are ravaged by an overactive immune response that sometimes manifests in severe cases of COVID-19. But such patients may still need a fully functioning immune system to fend off the virus itself.

The RECOVERY trial indicates that at the doses tested, the benefits of steroid treatment can outweigh the potential harm. The study found no outstanding adverse events from the treatment, investigators said. "This treatment can be given to pretty much anyone," says Horby.

And the pattern of response — with a greater impact on severe COVID-19 and no effect on mild infections — matches the notion that a hyperactive immune response is more likely to be harmful in long-term, serious infections, says Anthony Fauci, head of the US National Institute of Allergy and Infectious Diseases. "When you're so far advanced that you're on a ventilator, it's usually that you have an aberrant or hyperactive inflammatory response that contributes as much to the morbidity and mortality as any direct viral effect."

Easy to administer

So far, the only other drug shown to benefit people with COVID-19 in a large, randomized, controlled clinical trial is the antiviral drug remdesivir. Remdesivir shortened the amount of time that patients might need to spend in hospital, but it did not have a statistically significant effect on deaths (J. H. Beigel *et al*. *N. Engl. J. Med*. http://doi.org/dwkd ; 2020).

Remdesivir is also in short supply. Although the drug's maker, Gilead Sciences of Foster City, California, is ramping up production, the drug is currently available to only a limited number of hospitals around the world. And remdesivir is complex to administer: it must be given by injection over the course of several days.

Dexamethasone, in contrast, is a medical staple found on pharmacy shelves worldwide and is available as a pill — a particular benefit as coronavirus infections continue to rise in countries with limited access to health care. "For less than £50 (US$63), you can treat eight patients and save one life," says Martin Landray, an epidemiologist at the University of Oxford, and another chief investigator on the RECOVERY trial.

RICARDO CASTELAN CRUZ/EYEPIX GROUP/BARCROFT MEDIA VIA GETTY

Health workers in Mexico examine a person with COVID-19 who is on life support.

LATIN AMERICAN SCIENTISTS JOIN THE CORONAVIRUS VACCINE RACE

Researchers fear that breakthroughs from abroad will be too slow or inequitably shared to benefit the global south.

By Emiliano Rodríguez Mega

Gustavo Cabral de Miranda is used to people doubting him. As a child, he often had to put his schoolwork on hold to help his family, selling ice cream at fairs or working at a butcher's shop in northeastern Brazil. By the time he decided, at 22 years old, to study to become a scientist, others were telling him that academic life would not suit him: "It wasn't for people like me," he remembers them saying.

Now, Cabral, an immunologist at the University of São Paulo, is one of a number of ambitious Latin American scientists who are forging ahead with vaccine research programmes to fight COVID-19.

Right now, there is no vaccine for the coronavirus that causes the disease. A select group of candidates, most of them supported by pharmaceutical companies in China, the United States and Europe, have entered trials in humans. But researchers such as Cabral want a back-up plan, in case these well-resourced front runners are not successful, or hoarding or international deal-making prevents them from reaching low- and middle-income countries. Their goals echo long-standing efforts throughout Latin America to capitalize on national knowledge and establish — or re-establish — scientific independence from overseas pharmaceutical companies.

As Latin America becomes the new epicentre of COVID-19, concerns are flaring about the prospect of relying on a vaccine developed and manufactured elsewhere, especially given that rich countries have had better access to vaccines in the past. "We've already seen some monopoly behaviour, even though we don't have a COVID-19 vaccine yet," says Gavin Yamey, a global-health researcher at Duke University in Durham, North Carolina. Some governments of high-income countries have reportedly tried to buy vaccine-manufacturing companies or acquire part of their supply.

"The only ones who are going to solve the problems in Latin America are going to be us, Latin Americans. No one's coming to rescue us," says María Elena Bottazzi, a Honduran microbiologist at Baylor College of Medicine in Houston, Texas, who's developing a COVID-19 vaccine that she plans to distribute in countries throughout the region.

Some groups are working on ensuring equitable access, but billions of doses will be needed worldwide and no single provider can supply that amount, says Fernando Lobos, a

director at Sinergium Biotech, a vaccine maker in Buenos Aires.

Instead of waiting to see what happens, researchers across Latin America are working to find their own way out of the pandemic. "It does not matter if we start with less funds, but rather that we start," says Cabral.

Betting on innovation in Brazil

Cabral returned to Brazil in November 2019, after five years in Europe learning about new vaccine technologies. He was working on vaccines against the bacterium *Streptococcus pyogenes* and the Chikungunya virus, which both cause a lot of illness in Brazil, when COVID-19 began spreading rapidly across the globe. "I had to quickly adapt the project," he says.

His team uses harmless, hollow, virus-like particles created in the lab. The researchers stud the surfaces of these particles with fragments of the proteins that the coronavirus uses to enter human cells; the idea is to trick the immune system into producing antibodies that block the coronavirus.

The virus-like particles can't replicate in the human body, so they're considered safer than vaccines made from weakened viruses. Cabral's group is starting animal testing, and some other COVID-19 vaccine candidates in preclinical evaluation are following a similar approach.

Cabral thinks Brazil would be able to produce a safe and effective vaccine on a large scale and distribute it throughout the nation and to neighbouring countries. The country is one of the largest vaccine producers in Latin America. Luciana Leite, a vaccinologist at the Butantan Institute in São Paulo, says that innovation is key. The world will need alternatives if vaccine candidates using conventional approaches fail. "If you have an idea that's different from what's out there, I think it's worthwhile contributing with that," she says.

Her own approach is based on tiny bubbles, or vesicles, released by some bacteria to mislead a host's immune system. Researchers have tried to harness these vesicles to carry viral proteins — antigens that the immune system can recognize and make antibodies against — mixing them together to trigger a response. In as-yet-unpublished research, Leite and her students have found a way to attach large amounts of antigens to the bubbles, inducing a stronger-than-usual immune response.

Cabral hopes that all these efforts will show Brazilians that the government should invest more in research. The COVID-19 pandemic is the right moment, he says.

"This is the best time to open our eyes," Cabral says. "When the dust settles, priorities change."

Forging partnerships in Mexico

A few weeks ago, Laura Palomares reached out to a colleague for help with a COVID-19 vaccine she is developing using virus-like particles. Palomares, a biotechnologist at the National Autonomous University of Mexico in Cuernavaca, was shocked by his response.

"He says, 'Laura, all right, I'm going to help you, but I don't know why you're wasting your time doing this.'" Why bother creating a vaccine when the first successful vaccines will come from abroad, she remembers him asking. It's an argument she has heard many times — but it ignores Mexico's history of vaccine production.

For four decades, Mexico manufactured most of the vaccines needed for its national immunization programme. But the government dismantled the institutes responsible for producing vaccines and, in the late 1990s, replaced them with Birmex, a state-owned company. Since then, national vaccine production has plummeted. Now, Palomares says, Mexico produces only two vaccines, against influenza and hepatitis B.

If her vaccine works against SARS-CoV-2, she might be able to partner with Birmex to get it produced. Palomares hopes that manufacturing a domestically developed COVID-19 vaccine will push the company to resume the production of other vaccines needed for the national immunization programme.

But another option might be seeking help from abroad, says José Manuel Aguilar, a biotechnologist at the Monterrey Institute of Technology and Higher Education in Mexico. He and his colleagues have already contacted a Canadian company that could manufacture enough vaccine to start evaluating it in humans once they finish animal testing.

"The only ones who are going to solve the problems in Latin America are going to be us, Latin Americans. No one's coming to rescue us."

Their vaccine uses rings of DNA called nanoplasmids, which are taken up by cells. The DNA encodes a specific region of the protein that SARS-CoV-2 uses to infect a host; this allows cells to produce the antigen and display it, alerting the host's immune system.

Aguilar says that if the group's clinical results are promising, and if the researchers can find enough funds, they could ramp up production in three months, to fabricate millions of doses. "We could put Mexico in the vaccine race," he says.

Playing to strengths in Chile

In the 1960s, Chile developed a vaccine that soon became key to controlling rabies throughout Latin America. Since then, however, the country has lost its production capacity. Today, "all human vaccines used in Chile are obtained from foreign laboratories", says Alexis Kalergis, an immunologist at the Pontifical Catholic University of Chile in Santiago.

But that hasn't stopped Chilean researchers from trying to tame other viral diseases. Kalergis himself has become well known since he and his team developed a vaccine candidate that might one day protect newborn babies against the respiratory syncytial virus, a major cause of pneumonia and bronchitis in young children.

Now, his laboratory is rushing to create four prototype vaccines against SARS-CoV-2 using protein fragments, genetic material and live but harmless bacteria that can express some of the coronavirus's components. If everything goes to plan, studies in humans could begin as soon as next year. And if any of his prototypes shows promise, Kalergis says, he will reach out to vaccine-producing companies outside Chile to help manufacture it.

"Our plan is to distribute this vaccine against SARS-CoV-2 to all countries that need it," he says, "but with emphasis in our region."

A visit to the vet in Peru

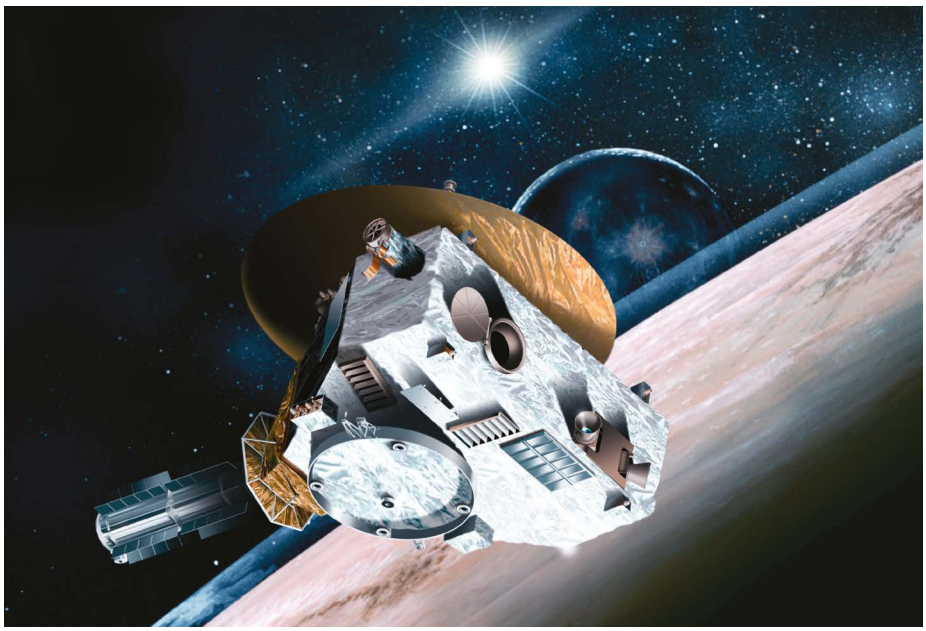
As cases of the coronavirus disease climbed in China in early January, Mirko Zimic had a hunch. "I had a very strong suspicion that it would cross borders and affect many nations, in particular Peru," he says. Zimic, a biophysicist at Cayetano Heredia University in Lima, had worked on vaccines for use in chickens and pigs. He called colleagues at FARVET, a veterinary pharmaceutical company in Chincha, and proposed that they work together on something neither had done before: making a human vaccine.

The researchers used insect cells to manufacture coronavirus spike proteins to provoke an immune response. Zimic and his colleagues will inject these into mice to test them. But there's an issue: if those tests are successful, the next step will be to immunize monkeys and then infect them with the coronavirus to see if the vaccine works. Peru has very little capacity to perform such experiments, Zimic says.

In 2016, FARVET began negotiations with Peru's science agency, CONCYTEC, to be allowed to manufacture human vaccines. But Zimic suspects that intense political turmoil stopped the deal; since 2016, three former Peruvian presidents have been accused of corruption. One killed himself before being detained.

Searching for a COVID-19 vaccine represents much more than a way out of the pandemic, says Zimic. It's an opportunity to make the Peruvian government aware of the need to value science. Some things could soon change: FARVET has again reached out to health authorities to be certified as a human-vaccine maker.

"We will not always be able to buy or import solutions," says Zimic. "It is my dream that in Peru, as well as in several countries in Latin America, we start producing our own vaccines."



JHUAPL/MSFC/NASA

The New Horizons craft (artist's impression) has captured images of Proxima Centauri.

PLUTO PROBE OFFERS FRESH VIEW OF SUN'S CLOSEST NEIGHBOUR

The unprecedented shots are set to be followed by other deep-space astronomy efforts.

By Davide Castelvecchi

NASA's New Horizons probe wowed the world in 2015 with unprecedented pictures of Pluto, and, more recently, with the first close-up images of an object in the Kuiper belt of asteroids. Now the mission has achieved yet another first: measuring the distances of two stars from the outer reaches of the Solar System.

"It's fair to say that New Horizons is looking at an alien sky, unlike what we see from Earth," said Alan Stern, the New Horizons principal investigator, at the Southwest Research Institute in Boulder, Colorado, in a statement released on 11 June.

Most space telescopes, from the venerable Hubble Space Telescope to the brand new European planet hunter CHEOPS, stay in the vicinity of Earth: being outside the atmosphere is enough to provide a great view, and there is usually no reason to venture farther afield. But there are sometimes advantages to making observations from deep space. Other spacecraft that will be used to do such deep-space astronomy include NASA's Curiosity rover. Astronomers will use it later this year to observe the star Betelgeuse, which started to mysteriously dim last year.

On 22 and 23 April, the New Horizons team pointed the probe's main camera at Proxima Centauri — the star closest to the Sun — which is about 1.3 parsecs away, as well as at another star, called Wolf 359. NASA asked professional and amateur astronomers to take pictures of these two stars, at exactly the same time, from Earth. Because New Horizons is now 46 times farther from the Sun than Earth is, the two points of view are distant enough for the stars' positions to look slightly different with respect to other, more distant objects. By measuring that difference, astronomers can calculate the two stars' distances from Earth.

This time-honoured technique, called parallax, is at the heart of the most sophisticated 3D maps of the Milky Way Galaxy, including the current state-of-the-art map, made by the European Space Agency's Gaia probe.

But Gaia is stationed relatively close to Earth, and it calculates parallax by comparing its views of the same stars six months apart, that is, from either side of half an orbit around the Sun. Those two positions differ by merely twice the Sun–Earth distance — not 46 times, as in the case of the New Horizons probe. As a consequence, the parallax angles are tiny, and Gaia produces tables of numbers rather than something for people to look at.

"For all the impressive work Gaia does, you can't see it," says New Horizons team member Tod Lauer, an astronomer at the US National Optical-Infrared Astronomy Research Laboratory in Tucson, Arizona. "Here, you can see it — KaBlam!" The two shots of Proxima Centauri, one from the Kuiper belt and the other from Earth, show the star clearly shifting position. The images could become as iconic and memorable as the celebrated Pale Blue Dot, a picture of Earth taken by NASA's Voyager 1 probe in 1990, Lauer says. The team has already been contacted by authors of astronomical textbooks who want to include the images in their next editions.

Watching from afar

Putting observatories into deep space — that is, anywhere beyond the Earth–Moon system — could offer a number of advantages. The Laser Interferometer Space Antenna (LISA), a trio of space probes that the European Space Agency plans to launch in 2034, will detect gravitational waves from a vantage point far from the disturbances of Earth. And this July, NASA's Curiosity rover will make a unique observation of the star Betelgeuse, in the constellation of Orion, from the surface of Mars. The star had a period of unusual dimness during the past year, temporarily losing more than two-thirds of its brightness. If Martian weather cooperates — the red planet is notorious for its dust storms — Curiosity will be able to see whether Betelgeuse is continuing to behave oddly.

In the case of stellar-distance measurements, an observatory with Gaia's sophistication could get much more precise parallaxes if it orbited far from Earth. For example, at five times father from the Sun than Earth is — Jupiter's distance — measurements would become five times more precise, at least in principle.

But astronomers have rarely considered sending probes into deep space. One reason is the time it would take for these instruments to gather good measurements. Gaia needs to orbit the Sun multiple times and to measure stars repeatedly to get a good parallax, and anything orbiting at Jovian distances would take much longer to do that, says Michael Perryman, an astronomer at University College Dublin. "Neglecting the energetic problems of getting a satellite there, and slowing it down, this would require a measurement duration of 3 to 5 orbits, or between 36 and 60 years," says Perryman. "Enough said!"

Because New Horizons is combining its shots with images taken from Earth, it can get a parallax in one go, without having to wait. The probe was never designed to do astronomy, so its measurement of Proxima Centauri's distance is orders of magnitude less precise than Gaia's, Lauer says. But he adds that getting a better measurement was never the goal. The point was to demonstrate how far human ingenuity has come.

PAVING OVER THE PAST

DR. Z. LEDERMAN, TEL BETH SHEMESH EXCAVATIONS

Israel is in the middle of a building boom to house its rapidly growing population, but some researchers fear the country isn't doing enough to conserve its wealth of archaeological sites. **By Josie Glausiusz**

The face carved in limestone is a vestige of a vanished world. With two dots for eyes and a slight hint of a smile, the 7,000-year-old figurine could be a ritual object, perhaps an amulet, or even a simple doll. The thumb-sized face is one of several dozen figures — mostly of goats and sheep — unearthed during an archaeological exploration lasting almost three years at En Esur in Israel[1], about 52 kilometres north of Tel Aviv.

The excavation at En Esur, also known by its Arabic name of Ein Asawir, "is an extraordinary project", says Dina Shalem, an archaeologist employed by the Israel Antiquities Authority (IAA), who co-directed the dig with IAA archaeologists Yitzhak Paz and Itai Elad. By the Early Bronze Age, 5,000 years ago, Paz says that En Esur was a "mega-city, the largest so far known in the Southern Levant", a region spanning modern Israel, the Palestinian territories and Jordan. Excavating En Esur was, he says, "a once-in-a-lifetime experience".

Built over the remains of an earlier, smaller village (from which the stone face was unearthed), the metropolis spanned an estimated 65 hectares and was home to between 5,000 and 6,000 people; more than 20 times the typical size of villages in that area at the time. Thanks to a year-round flowing spring, the townspeople of En Esur thrived, growing wheat, barley, lentils, grapes and olives, and raising cows, pigs, sheep and goats.

A visit to the site in November 2019 during an excavation showed how enormous the place once was. Stretching into the distance were the remains of house foundations and alleyways. A grand, 600-square-metre temple enclosed two massive stone basins — the larger of which was 3.3 metres long and was filled with burnt animal bones, possibly from sacrifices. "We were really amazed at how densely built the city was," Shalem says, "the planning, the streets". A gigantic pile of approximately 5 million pottery shards, excavated from the site, attests to the domestic life of this bustling town. "Pottery, flint, figurines, burials — we can tell that it's a complex society," she says.

There's a lot for the archaeologists and labourers to label and store for shipment. But they were not the only ones working at the site. Engineers were also taking measurements for Netivei Israel, the country's transport infrastructure company, which funded the archaeological excavation in preparation for the building of a road intersection on part of the site. This vanished world, briefly uncovered,

Road building threatens the site of Tel Beit Shemesh, dating to at least the seventh century BC.

disappeared again from view when it was covered with earth and cement over the winter.

En Esur is a huge site, so most of it is still underground and untouched. But the massive temple and the other excavated parts will remain buried under the road intersection for decades – possibly longer. Critics charge that this important evidence will never be seen again.

> **“There are no red lines or any understanding or rules [for] what should be kept.”**

This year, hundreds of other archaeological sites might also be buried or destroyed. That’s because most excavations in the country are salvage digs, authorized by the IAA, the government body that oversees antiquities and archaeological sites within the state of Israel. Salvage digs are conducted to document archaeological remains in danger of destruction because of development plans. But the IAA very rarely blocks construction on top of an important archaeological site or takes steps to preserve some portion of the site from being destroyed, says Yonathan Mizrachi, chief executive of Emek Shaveh, an Israeli non-governmental organization based in Jerusalem that works to protect ancient sites as public assets.

In 2019, salvage digs accounted for more than half of the 424 licences issued by the IAA for archaeological excavations and surveys, according to the agency. In almost all cases, after the archaeologists excavated the sites and removed valuable artefacts, construction projects were allowed to proceed, says Mizrachi.

Paving over archaeologically valuable sites isn’t unique to Israel, Mizrachi says. In Turkey, for example, the 12,000-year-old town of Hasankeyf, a monumental site on the Tigris river, has slowly been submerged beneath the new Ilısu dam. But some archaeologists say that the situation is particularly problematic in Israel compared with many other countries. One big issue is that the majority of the IAA’s budget comes from the salvage digs before construction projects – and the government itself is the biggest developer in Israel, says archaeologist Uzi Dahari, a former deputy director of the IAA. The country is also building rapidly to keep up with a surging population. Archaeologists also argue that the Israeli government favours saving ancient Jewish sites, especially in Jerusalem, over ones linked to other religions.

In Israel, says Mizrachi, “there are no red lines, or any understanding or rules [for] what should be kept and should be destroyed”.

Gideon Avni, head of the archaeology division at the IAA, says that rather than block construction after a salvage dig, the IAA tries to prevent development projects in advance if a site is known to be important. The agency, he says, follows strict criteria established by the United Nations Educational, Scientific and Cultural Organization when deciding whether or not a site should be permanently saved, or covered and built on. He adds that “every act that we are doing is subject to public monitoring, professional monitoring, our internal system, which tries to prevent this conflict of interest.”

And some archaeologists in Israel say that many sites would not be excavated at all if it weren’t for the construction projects that provide funding and an impetus for digs. Given Israel’s fast-growing population, they say, the country can’t preserve as many archaeological sites as supporters would like. “I would like the whole country to be covered in archaeology,” says Avni. “But my children need a place to live.”

Growing pressure

Compared with other countries around the Mediterranean, Israel has a much higher concentration of archaeological sites in a smaller area, Dahari says. There are about 35,000 sites in a country of 22,145 square kilometres.

An average of 200–300 archaeological sites are excavated each year as rescue or salvage digs before development. But in evaluating whether construction at the sites can go ahead, the IAA has a conflict of interest because it is heavily funded by the construction industry, says Dahari.

In 2019, construction-funded salvage digs accounted for 83% of the IAA’s budget of

426 million shekels (US$122.7 million). "On the one hand, their responsibility is to restore and to protect archaeological sites, but on the other hand, if you will tell the constructor, 'don't build, because this is an archaeological site', where will the money come from for the IAA? That's a big contradiction," Dahari says. "I don't blame the Israel Antiquities Authority," he says, "but I blame the government, because of the budgetary system of salvage excavations."

Yorke Rowan, an anthropological archaeologist at the University of Chicago, Illinois, agrees. "That is one contradictory aspect of the way the IAA works: a major bureaucracy depends on funding from the construction industry to fund itself." Rowan, who has studied stone bowls, mortars and grinders dating to 5000–3000 BC from previous digs at En Esur, says the difficulty is that the IAA is "mandated to preserve and protect archaeological sites and heritage, but that funding is tied to positive outcomes for development projects".

Adding to the pressure is Israel's rapidly growing population. The nation had a fertility rate of 3.1 children per woman in 2018, compared with 1.7 for the United States and an average of 1.6 in the member countries of the Organisation for Economic Co-operation and Development, to which Israel belongs (see 'Population boom'). Israel's current population is more than 9 million; by 2065, according to the Israel Central Bureau of Statistics, the population is predicted to swell to 20 million, which would make the country one of the most densely populated on the planet.

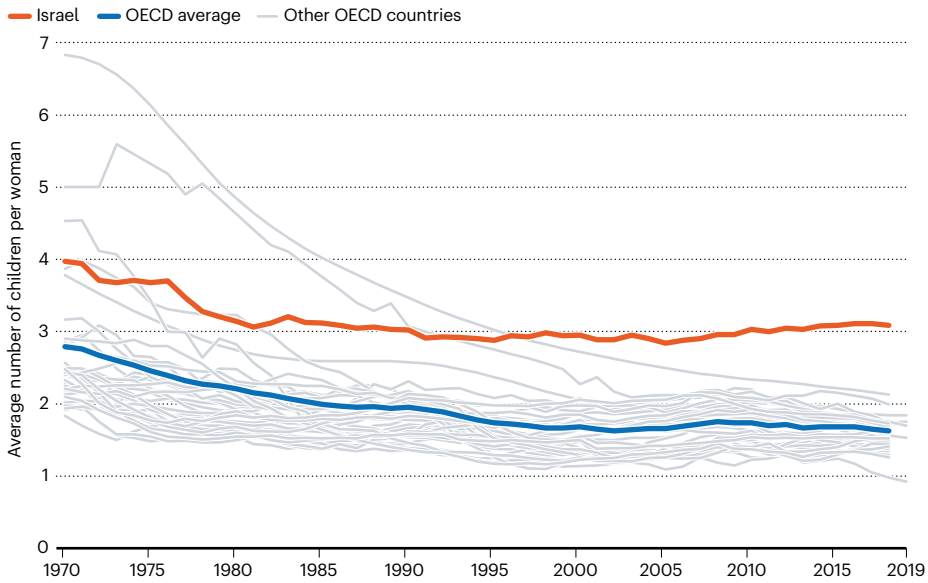
The effects of this population boom are obvious across the country. New neighbourhoods crammed with high-rise apartment blocks are springing up on the outskirts of many towns; multi-lane highways criss-cross the country; gleaming malls and industrial zones mark the landscape. Between 2014 and 2017, according to HaMaarag, a consortium of environmental agencies based in Jerusalem, Israel lost 107 square kilometres of undeveloped land to construction and farming.

With population growth driving so much construction, says Dahari, "the IAA cannot say, 'this is an archaeological site, don't build anything' – that's impossible". He says the authority "is not finding the middle way between the needs of development and the protection of the archaeological site". Dahari says that during his tenure at the IAA, the agency classified about 700 of the 35,000 sites in Israel as unique archaeological sites and therefore off-limits for development. But some of these 700 sites "are only partially protected", he says.

The list includes a site called Tel Beit Shemesh, which dates back at least to the seventh century BC. The settlement was possibly a large-scale centre for olive-oil production[2]. Archaeologists regard this site as important because it was located between two rival population centres more than two and a half millennia ago. It would take archaeologists decades to study this kind of site, but a four-lane highway will soon run through it. That is, says Dahari, "in my point of view, a crime against archaeology". Avni acknowledges that the construction at Tel Beit Shemesh, "was a very, very painful compromise for us", balancing the need to preserve the site and building a highway that would serve up to 250,000 people. He said the current plan is to recreate the site in a park on a bridge over the highway.

POPULATION BOOM

Israel's fertility rate is the highest in the 37 countries in the Organisation for Economic Co-operation and Development (OECD).



SOURCE: OECD

National narratives

Time is also running out for an ancient hill village called Nebi Zechariah that is at least 2,000 years old. The excavations of the village lie in an industrial zone on the outskirts of Modiin, a city about 26 kilometres northwest of Jerusalem. The excavated buildings have a forsaken air. The stone walls, doorways and mosaic-tiled floors are overgrown with brambles and weeds; a nearby billboard trumpeting a new industrial and logistics centre is pasted with a red sign saying "sales closed". In the distance, the high-rise blocks of the expanding city of Modiin can be glimpsed. By 2040, according to the city plan, the addition of 43,000 housing units will enable the population to more than double, from 93,000 to 240,000. And those plans call for construction in the city's industrial zone, including over Nebi Zechariah.

The ancient site was clearly once a bustling place. Originally founded as a Jewish village during the Roman period, beginning in 63 BC, it was occupied continuously for about 1,000 years by waves of polytheists, Byzantine Christians and Muslims, living together until it was abandoned in the eleventh century during a period of climate-change-induced drought. Its inhabitants built luxurious houses paved with mosaics, chiselled Christian crosses and Greek inscriptions into their olive presses, and produced glass weights inscribed with Arabic script for weighing coins.

"There is a very interesting story in this type of site, because it consists of a kind of continuity of the rural population of ancient Palestine from Roman until Crusader times," says Avni. "It also shows a kind of living together of different communities in a settled place at the same time."

But that does not mean the decision to build on it was wrong, according to Avni. He says that the settlement is just one out of 400 similar sites of the period in Israel. "It's not unique when you look at the whole country," he says. And population pressures inevitably lead to compromises that enable modern development. Even so, "it's really a pity, when you look at it as a citizen, to destroy it", he says. "It's a beautiful site, and could have been preserved as a park."

That's unlikely to happen. Eyal Malul, a spokesperson for the municipality of Modiin, says part of the land is designated for building.

Nebi Zechariah and sites like it might have a better chance of being preserved if they dovetailed with Israel's national narrative, Mizrachi suggests. "In general, the Israeli government, for generations, is dealing with Jewish history," he says. As an example of the government's interest in preserving Jewish history through archaeology, Mizrachi and other researchers point to a site called the City of David, located in the Palestinian neighbourhood of Silwan in East Jerusalem. An organization called the Ir David Foundation (El-Ad) is authorized by the government to sponsor excavations there and run the City of David National Park. El-Ad claims that the biblical King David built a palace in the City of David 3,000 years ago and says it is "dedicated to the preservation and development of the Biblical City of David".

But there is serious disagreement among archaeologists as to whether the large structure that El-Ad identified as King David's palace was actually from the era in which the biblical king supposedly lived, and whether this ancient stone structure can be linked to him. "El-Ad's narrative is based on biblical history, not archaeology. They are content as long as archaeology doesn't contradict their reading," says Raphael Greenberg, an archaeologist at Tel Aviv University. He argues that El-Ad is "supported by the Israeli administration, that uses archaeology selectively to market its ideology".

El-Ad rejects that argument. Doron Spielman, a spokesperson for the organization, says that there have been 20 excavations of the site over 150 years and they have found numerous carved inscriptions and clay stamps with the names of biblical figures. "There is no other place on Earth that has found more corroborative evidence of the biblical story than the City of David."

When it comes to saving or excavating ruins, Avni says, "we have been accused many times of giving preference to Jewish sites or synagogues. When you look at the history of archaeology in this country, it's not baseless." But he says that those criticisms applied to the early years of the country's modern history, and that now, "most of the sites we are excavating are Byzantine Christian, early Muslim, medieval Muslim and Ottoman." Avni's own speciality is early Islamic archaeology.

Yet some researchers say that when archaeological sites involve religions other than Judaism, they sometimes get less protection. And many Israelis aren't aware that people of different religions coexisted in the region after Islam arrived in the seventh century AD, says Mizrachi. He adds that sites from the early Islamic period "are extremely important to be preserved and Nebi Zechariah is a very good example of that".

> "What we rescue is the knowledge, not necessarily the site."

Possible solutions

Without financing, however, the likelihood of preserving such sites is slim. Unlike the excavations funded by El-Ad, very few of the digs that the IAA conducts are financed independently. There would be no excavation at En Esur had a road junction not been planned there, and laboratory research on material collected at the site will continue for years, say the IAA archaeologists who work there. "What we rescue is the knowledge, not necessarily the site," says Paz. He and the other IAA archaeologists who excavated En Esur declined to say whether the site itself should have been preserved.

But others aren't so guarded. Greenberg, an expert on the Bronze Age in the Levant and a former IAA employee, questions the speed at which the salvage dig at En Esur was conducted, although he emphasizes that this is not the fault of the archaeologists but of the system that forces rapid excavations. "Because the site was excavated under severe time constraints, and because of the unprecedented quantity of finds that must be processed, it will be years before we have a detailed understanding of the results," he says.

Less-destructive alternatives exist, Greenberg contends. For example, he suggests that 5% of the threatened site could have been set aside and excavated thoroughly, with the road project covering the rest of it.

Greenberg has general concerns about speedy salvage excavations. Academic excavations typically take many years. Researchers alternate time in the field with longer stretches in the lab to study findings and adjust their plans for later stages of the dig. That's not possible during a salvage operation, he says.

The kind of alternative that Greenberg envisaged for En Esur is sometimes possible in other countries, such as the United States and France, even if excavations do get paved over there, too. Archaeologist Morag Kersel at DePaul University in Chicago, points to the Miami Circle in Florida, a round structure linked to the Native American Tequesta people that was scheduled to be destroyed, but was preserved after protests. Sometimes, she says, "when it is feasible, projects are redesigned to avoid the destruction of the site".

One difference in the United States, explains Rowan, is that "there is an extra level of oversight at the state level". Each US state has a State Historic Preservation Office that oversees surveys and excavations. If nothing important is found during a salvage survey, the office will approve the site for development, but it can mandate extra testing, including remote sensing and excavation, if significant evidence is found.

"Something like En Esur, if it was in the United States, that would be stunning. There's not a chance you would pave it over," says Rowan. "I have great faith that they did a good job, because I know the archaeologists. It does seem a shame to pave over a site like that."

Archaeologist Pierre de Miroschedji, former director of research at the French National Centre for Scientific Research in Nanterre, calls En Esur "a great discovery". In France, he says, "it has happened many times that the line of a highway had to be changed because an important discovery was made".

Even so, his approach is pragmatic: "We have to look for an equilibrium, for balance, between the necessity of modern life and the necessity of preserving the heritage."

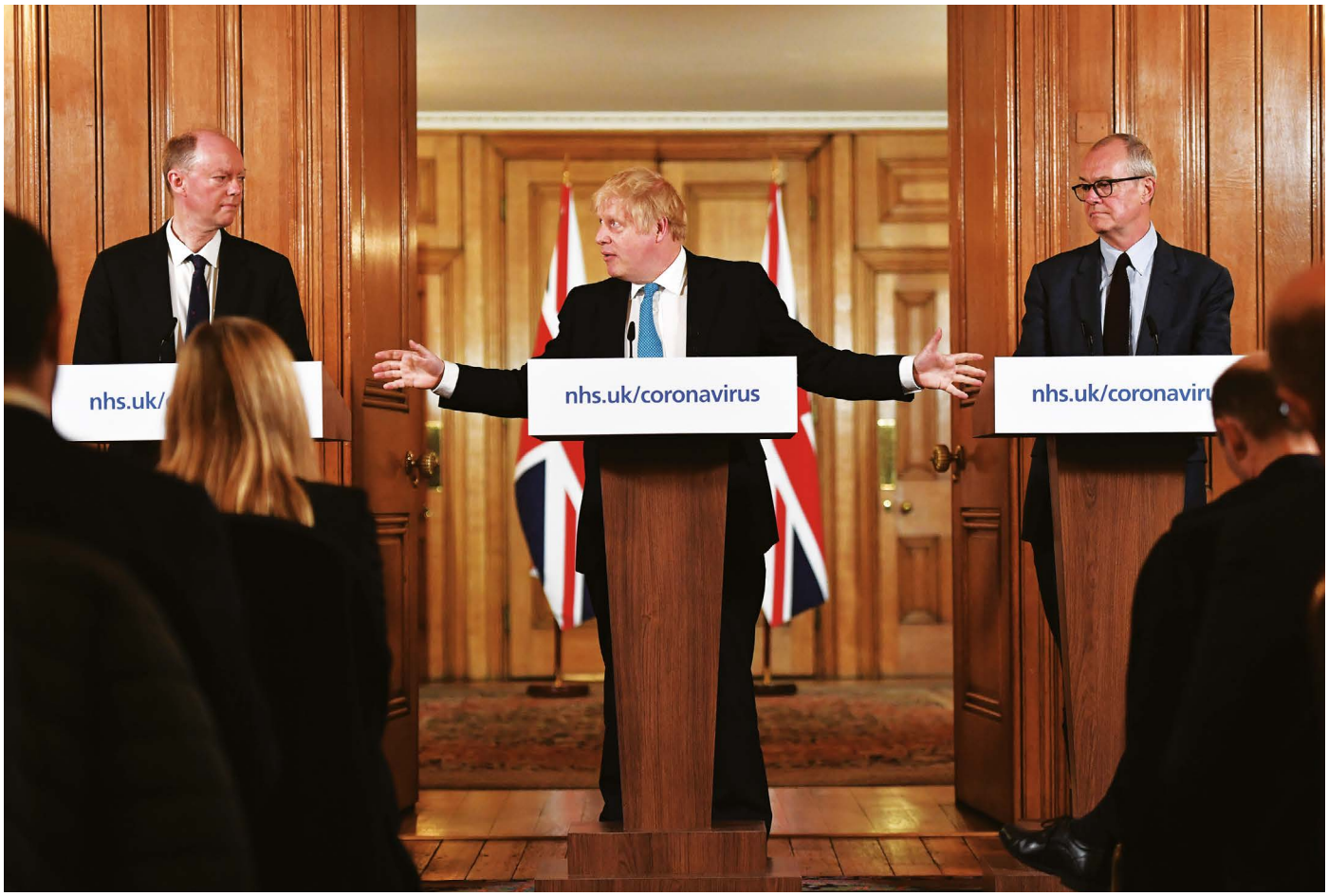
HEIDI LEVINE FOR NATURE

Archaeologist Dina Shalem (left) with a clay vessel from the ancient village of En Esur.

Josie Glausiusz is a science journalist in Israel.

1. Elad, I. & Paz, Y. *'En Esur ('Asawir). Excavations and Surveys in Israel* Vol. 130 (Israel Antiquities Authority, 2018).
2. Haddad, E. & Ben-Ari, N. *Bet Shemesh, Tel Bet Shemesh. Excavations and Surveys in Israel* Vol. 131 (Israel Antiquities Authority, 2019).

Books & arts



LEON NEAL/BLOOMBERG VIA GETTY

UK Prime Minister Boris Johnson (centre), with chief medical and scientific advisers at the daily press conference on COVID-19 in early March.

Scathing COVID-19 book from *Lancet* editor — rushed but useful

Richard Horton skewers leaders in two of the richest, most powerful and scientifically advanced countries for getting it so wrong. **By Stephen Buranyi**

Since the coronavirus crisis began, Richard Horton, editor-in-chief of leading medical journal *The Lancet*, has been tearing across the British public sphere. Here he is on the BBC, the national broadcaster, there in the pages of *The Guardian* newspaper — taking the government to task for failures that have left the United Kingdom with the world's second-highest per capita COVID-19 death toll so far (Belgium is top). Horton has never shied away from controversy (his journal published the retracted, fraudulent paper by Andrew Wakefield that alleged a non-existent link between vaccines with autism) or crusades (against the Iraq war and for political action on climate change). In coronavirus, he has found a cause that matches his energy: the Lancet journals are pumping out both the latest research and his pointed critiques of government policy; and last month, he reviewed a new book by the Slovenian Marxist philosopher Slavoj Žižek that imagines economic and social worlds after COVID-19.

Now Horton has a book of his own. *The COVID-19 Catastrophe* is a sort of history, diagnosis and prescription, in real time. It is wide ranging, querying the changing role of international cooperation and the fallout of austerity economics, and taking a deeper dive into China's scientific and political response to the crisis than most Western media have offered. But the book returns again and again

to the catastrophe in both the United Kingdom and the United States. It is haunted by the question: how did two of the richest, most powerful and most scientifically advanced countries in the world get it so wrong, and cause such ongoing pain for their citizens?

The easy answer is in their leadership. Horton levels the accusation that US President Donald Trump is committing a "crime against humanity" for defunding the very World Health Organization that is trying to help the United States and others. UK Prime Minister Boris Johnson, in Horton's view, either lied or committed misconduct in telling the public that the government was well prepared for the pandemic. In fact, the UK government abandoned the world-standard advice to test, trace and isolate in March, with no explanation, then scrambled to ramp up testing in April, but repeatedly failed to meet its own targets, lagging weeks behind the rest of the world. A BBC investigation in April showed that the UK government failed to stockpile neccessary personal protective equipment for years before the crisis, and should have been aware that the National Health Service wasn't adequately prepared.

Politicians are easy targets, though. Horton goes further, to suggest that although scientists in general have performed admirably, many of those advising the government directly contributed to what he calls "the greatest science policy failure for a generation".

Again using the United Kingdom as an example, he suggests that researchers were insufficiently informed or understanding of the crisis unfolding in China, and were too insular to speak to Chinese scientists directly. The model for action at times seemed to be influenza, a drastic underestimation of the true threat of the new coronavirus. Worse, as the UK government's response went off the rails in March, ostensibly independent scientists would "speak with one voice in support of government policy", keeping up the facade that the country was doing well. In Horton's view, this is a corruption of science policymaking at every level. Individuals failed in their responsibility to procure the best scientific advice, he contends; and the advisory regime was too close to — and in sync with — the political actors who were making decisions. "Advisors became the public relations wing of a government that had failed its people," he concludes.

The situation deteriorated to the point that former UK chief government scientific adviser David King set up an alternative scientific-advice committee on 3 May to both be more transparent to the public and exercise the independence of mind and tongue that he felt the official committees had ceded. It is a move Horton approves. It's not clear that the group has swayed government policy, but it has given the public and the media a stable and open source of scientific information.

ARUN CHANDRABOSE/AFP VIA GETTY

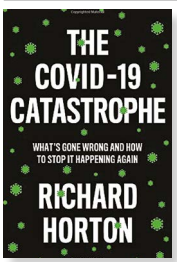
A 'walk-in sample kiosk' in Kerala, India, to test for SARS-CoV-2.

Elsewhere, the book discusses places with less-egregious failures, such as Spain and France, and the rare successes, such as New Zealand, and Kerala in India. Horton is fond of a tight timeline, and after experiencing the past six months through the fragmented lenses of social media and short news reports, it is a minor, but not insignificant, relief just to have all these events collected in one place, in order.

Such synthesis is a bulwark, too, against the blizzard of misinformation — the 'infodemic' — that has blanketed the crisis. Horton rightly points out that it isn't just quacks and conspiracy theorists promulgating false information. There is a chilling effort by governments as well. Examples include the United Kingdom's bizarre recent assertions that it never pursued a 'herd immunity' strategy, despite both the prime minister and a scientific adviser discussing it in public, and the Chinese authorities' gagging of physician Li Wenliang, who tried to raise the alarm at the very beginning of the coronavirus outbreak in Wuhan.

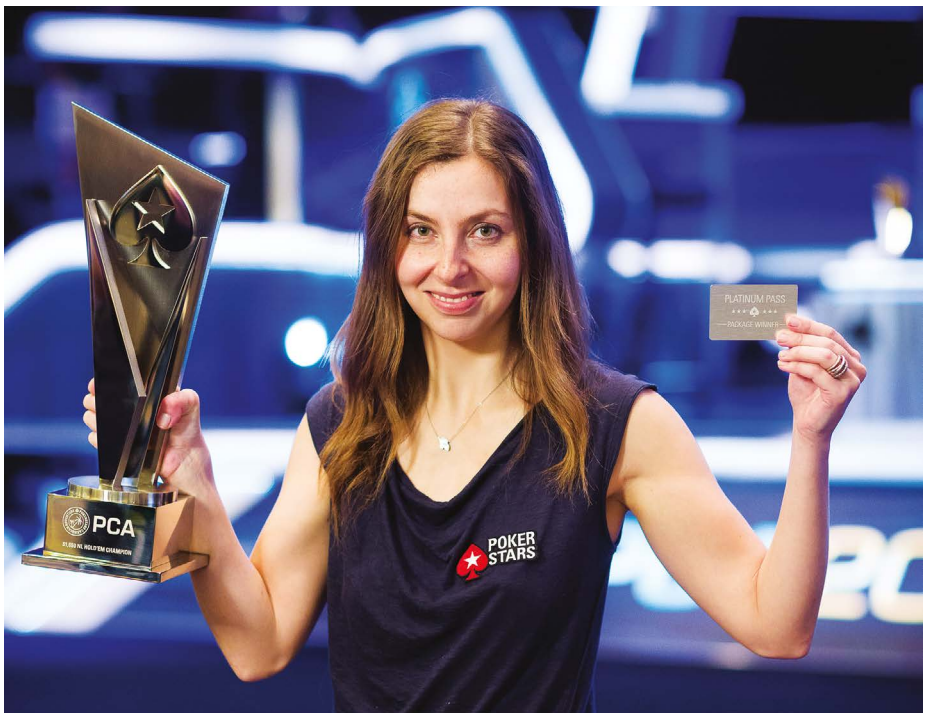
The book is weakest where it discusses the future. This is unsurprising — no one knows how this calamity will end. Horton is of the camp that believes there is unlikely to be a return to normality. "Perhaps Covid-19 represents an impermeable boundary between one moment of our lives and another. We can never go back," he writes. His recommendations for the future are largely admirable — a more-connected and conscientious public, a more-communitarian government, a more self-critical scientific elite. But there is little structure or path for getting there. I, too, want the United States to be a better global citizen, without a president such as Donald Trump. Sadly, I struggle to imagine how it will happen.

Like much of the scientific work produced during the pandemic, this book is a rushed product — not subject to the exacting standards of more normal times. But like the scientific work, it is vital and up to the minute. Early on, Horton paraphrases the mathematician and writer Adam Kucharski, saying "if you've seen one pandemic, you've seen … one pandemic." What is imparted now won't necessarily help us next time, and we're still in the swell of this one. The events the book recounts are barely cold, and the powerful whom it critiques are still in power. It has lessons that are useful right now.



The COVID-19 Catastrophe: What's Gone Wrong and How to Stop It Happening Again
Richard Horton
Polity (2020)

Stephen Buranyi is a writer specializing in science and the environment, based in London.
e-mail: stephen.gyorgy@gmail.com



Maria Konnikova won the 2018 PokerStars Caribbean Adventure National Championship.

NEIL STODDART/POKERSTARS

What the world needs now: lessons from a poker player

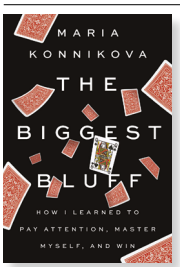
One scientist's zero-to-hero journey into risk, uncertainty and delusional models. **By Liv Boeree**

If the mess of public confusion and poor leadership surrounding the coronavirus pandemic has taught us anything, it is how poorly equipped we are to navigate risk and uncertainty. No crisis in recent memory has better showcased our systemic educational failings. In my nation – the United Kingdom – the population is mostly unable to think probabilistically, and the media establishment thrives on certainty and punishes humility around knowledge. It's a familiar picture elsewhere in the world, too.

There has never been a more pressing need for digestible and coherent literature on rational decision-making. Enter *The Biggest Bluff*, psychologist Maria Konnikova's depiction of her journey into professional poker. What at first seems a light-hearted story about a curious academic dipping her toe into shark-infested waters delivers a crucial lesson in how to thrive in an increasingly misleading world.

Konnikova finds poker when a run of horrible luck in her family sends her on a mission to understand the nature of chance. That introduces her to the work of John von Neumann – the founder of game theory. He loved poker for its practical, real-life applications, despite being a lousy player by all accounts. The game sits in a Goldilocks zone between the crisp, perfect information of chess (no hidden knowledge; best player almost always wins) and the mindless gamble of a roulette wheel. It involves just enough luck and just enough skill to resemble the messiness of reality.

As an astrophysics graduate who spent a decade playing poker professionally, I echo this sentiment. The game is a delicious mix of science and art. As Konnikova details, it demands quantified analysis and qualitative judgements, stress-testing players' reasoning in complex, high-pressure situations. It is a game of self-control and self-awareness, with a heap of Bayesian model-building thrown in.



The Biggest Bluff: How I Learned to Pay Attention, Master Myself, and Win
Maria Konnikova
The Penguin Press (2020)

And as *Nature* readers appreciate now more than ever, building models that extract meaningful signals from noisy data is tricky. Too specific and you might miss something key; too sensitive and your findings fail to replicate. The same is true in poker – you build mental models about each opponent and situation, with the added twist that the data are actively trying to deceive. "Why did his nostrils flare when that King hit?" "Why did he bet 80% of the pot instead of the usual 50%?" "Is he smiling because he's happy, or just pretending to be? Or is it just because I'm the only woman in the room?" Each hand provides a barrage of information in which you must sort fact from fiction.

As in life, the most perfectly crafted models and strategies can fall apart when the going gets tough. Mindset is key, and Konnikova wisely dedicates much of her book to this. Drawing on her expertise as a psychologist and the wisdom of her poker mentor Erik Seidel (one of the game's legends), she details what separates the best from the rest in a way reminiscent of Josh Waitzkin's 2007 chess-for-life book *The Art of Learning*. For example, she advises that we evaluate success by the quality of our mental process during a situation, not just by the outcome.

Where chess lacks the randomness of real life, poker shows how easily randomness can delude. Winning a tournament requires plenty of skill, but an awful lot of luck, too – an inconvenient truth that it is tempting to downplay. I learnt this the hard way after my European Poker Tour victory, a success that bred a dangerous complacency in me. Why would a freshly crowned champion need to keep working hard on her game? Whether you're a poker player or a scientist (or both), remember: the ego tends to ignore the luck factor in success.

A run of bad luck can be just as dangerous, demolishing self-confidence and making us change tactics when we don't need to. Despite her clear aptitude, some crushing early losses almost see Konnikova abandon the game entirely. She helps the reader avoid the same pitfalls by sharing her mistakes and moments of self-doubt with stark honesty.

As someone who has read almost every piece of literature on poker, I can say that *The Biggest Bluff* is the best depiction yet of the game I love, and the invaluable thinking skills it teaches. This is not a book that will teach you how to play your Ace–Queen out of position against a laggy villain (although it will teach you what that means). But it will show you how to play the game of life more effectively. Konnikova's is an uplifting zero-to-hero journey that will raise a smile in these trying times.

Liv Boeree is a science communicator and former professional poker player.
e-mail: info@livboeree.com

Comment

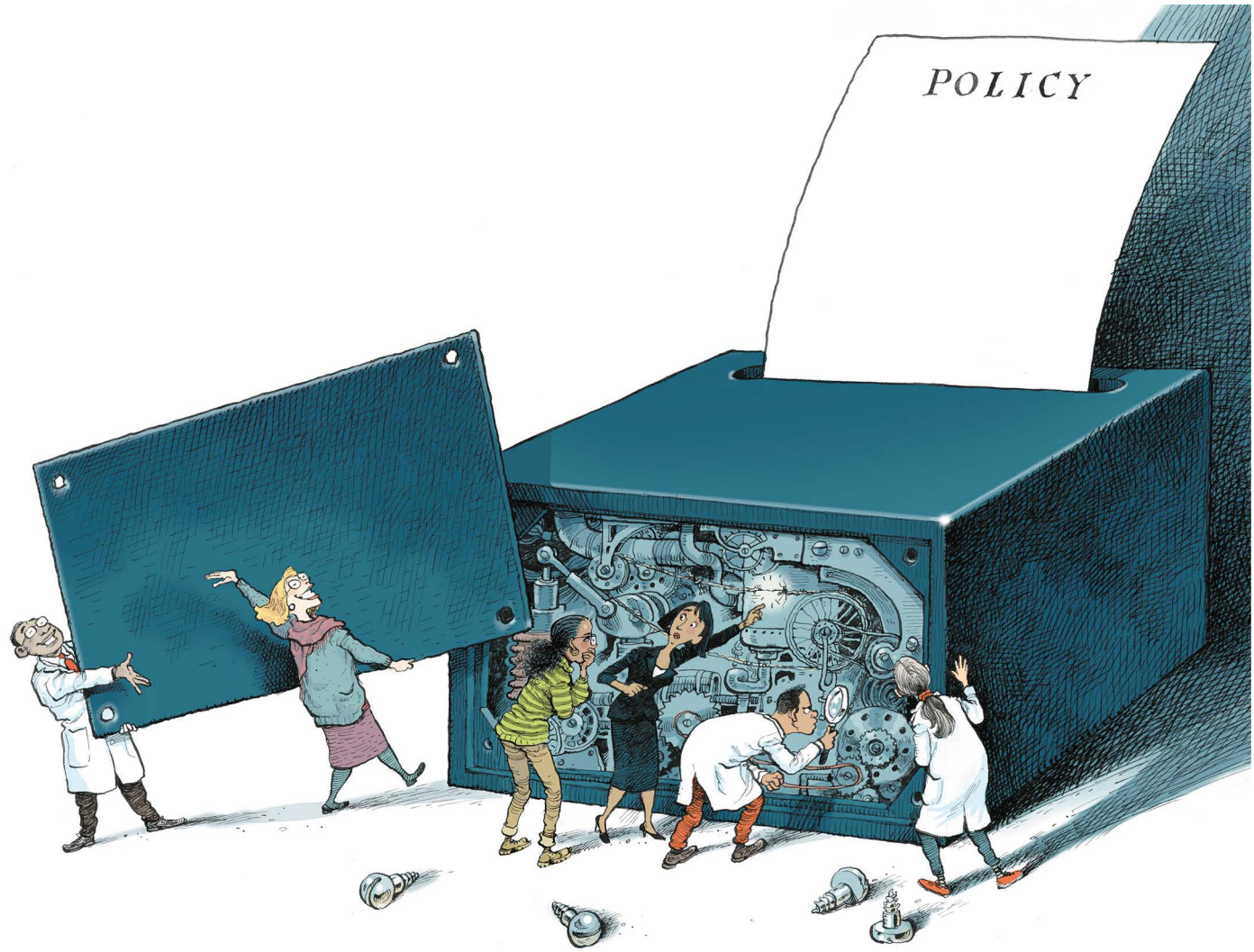


ILLUSTRATION BY DAVID PARKINS

Five ways to ensure that models serve society: a manifesto

Andrea Saltelli, Gabriele Bammer, Isabelle Bruno, Erica Charters, Monica Di Fiore, Emmanuel Didier, Wendy Nelson Espeland, John Kay, Samuele Lo Piano, Deborah Mayo, Roger Pielke Jr, Tommaso Portaluri, Theodore M. Porter, Arnald Puy, Ismael Rafols, Jerome R. Ravetz, Erik Reinert, Daniel Sarewitz, Philip B. Stark, Andrew Stirling, Jeroen van der Sluijs & Paolo Vineis

Pandemic politics highlight how predictions need to be transparent and humble to invite insight, not blame.

The COVID-19 pandemic illustrates perfectly how the operation of science changes when questions of urgency, stakes, values and uncertainty collide — in the 'post-normal' regime.

Well before the coronavirus pandemic, statisticians were debating how to prevent malpractice such as *p*-hacking, particularly when it could influence policy[1]. Now, computer modelling is in the limelight, with politicians presenting their policies as dictated by 'science'[2]. Yet there is no substantial aspect of this pandemic for which any researcher can currently provide precise, reliable numbers. Known unknowns include the prevalence and fatality and reproduction rates of the virus in

populations. There are few estimates of the number of asymptomatic infections, and they are highly variable. We know even less about the seasonality of infections and how immunity works, not to mention the impact of social-distancing interventions in diverse, complex societies.

Mathematical models produce highly uncertain numbers that predict future infections, hospitalizations and deaths under various scenarios. Rather than using models to inform their understanding, political rivals often brandish them to support predetermined agendas. To make sure predictions do not become adjuncts to a political cause, modellers, decision makers and citizens need to establish new social norms. Modellers must not be permitted to project more certainty than their models deserve; and politicians must not be allowed to offload accountability to models of their choosing[2,3].

This is important because, when used appropriately, models serve society extremely well: perhaps the best known are those used in weather forecasting. These models have been honed by testing millions of forecasts against reality. So, too, have ways to communicate results to diverse users, from the Digital Marine Weather Dissemination System for ocean-going vessels to the hourly forecasts accumulated by weather.com. Picnickers, airline executives and fishers alike understand both that the modelling outputs are fundamentally uncertain, and how to factor the predictions into decisions.

Here we present a manifesto for best practices for responsible mathematical modelling. Many groups before us have described the best ways to apply modelling insights to policies, including for diseases[4] (see also Supplementary information). We distil five simple principles to help society demand the quality it needs from modelling.

Mind the assumptions

Assess uncertainty and sensitivity. Models are often imported from other applications, ignoring how assumptions that are reasonable in one situation can become nonsensical in another. Models that work for civil nuclear risk might not adequately assess seismic risk. Another lapse occurs when models require input values for which there is no reliable information. For example, there is a model used in the United Kingdom to guide transport policy that depends on a guess for how many passengers will travel in each car three decades from now[5].

One way to mitigate these issues is to perform global uncertainty and sensitivity analyses. In practice, that means allowing all that is uncertain — variables, mathematical relationships and boundary conditions — to vary simultaneously as runs of the model produce its range of predictions. This often reveals that the uncertainty in predictions is substantially larger than originally asserted. For example, an analysis by three of us (A.Saltelli, A.P., S.L.P.) suggests that estimates of how much land will be irrigated for future crops varies more than fivefold when extant models properly integrate uncertainties on future population growth rates, spread of irrigated areas and the mathematical relationship between the two[6].

"The best way to keep models from hiding their assumptions, including political leanings, is a set of social norms."

However, these global uncertainty and sensitivity analyses are often not done. Anyone turning to a model for insight should demand that such analyses be conducted, and their results be described adequately and made accessible.

Mind the hubris

Complexity can be the enemy of relevance. Most modellers are aware that there is a trade-off between the usefulness of a model and the breadth it tries to capture. But many are seduced by the idea of adding complexity in an attempt to capture reality more accurately. As modellers incorporate more phenomena, a model might fit better to the training data, but at a cost. Its predictions typically become less accurate. As more parameters are added, the uncertainty builds up (the uncertainty cascade effect), and the error could increase to the point at which predictions become useless.

The complexity of a model is not always an indicator of how well it captures the important features. In the case of HIV infection, a simpler model that focuses on promiscuity turned out to be more reliable than a more involved one based on frequency of sexual activity[5]. The discovery of the existence of 'superspreading events' and 'superspreader' people with COVID-19 similarly shows how an unanticipated feature of transmission can surprise the analyst.

One extreme example of excess complexity is a model used by the US Department of Energy to evaluate risk in disposing of radioactive waste at the Yucca Mountain repository. Called the total system performance assessment, it comprised 286 sub-models with thousands of parameters. Regulators tasked it with predicting "one million years" of safety. Yet a single key variable — the time needed for water to percolate down to the underground repository level — was uncertain by three orders of magnitude, rendering the size of the model irrelevant[7].

Complexity is too often seen as an end in itself. Instead, the goal must be finding the optimum balance with error.

What's more, people trained in building models are often not drilled or incentivized for such analyses. Whereas an engineer is called to task if a bridge falls, other models tend to be developed with large teams and use such complex feedback loops that no one can be held accountable if the predictions are catastrophically wrong.

Mind the framing

Match purpose and context. Results from models will at least partly reflect the interests, disciplinary orientations and biases of the developers. No one model can serve all purposes.

Modellers know that the choice of tools will influence, and could even determine, the outcome of the analysis, so the technique is never neutral. For example, the GENESIS model of shoreline erosion was used by the US Army Corps of Engineers to support cost–benefit assessments for beach preservation projects. The cost–benefit model could not predict realistically the mechanisms of beach erosion by waves or the effectiveness of beach replenishment by human intervention. It could be easily manipulated to boost evidence that certain coastal-engineering projects would be beneficial[7]. A fairer assessment would have considered how extreme storm events dominate in erosion processes.

Shared approaches to assessing quality need to be accompanied by a shared commitment to transparency. Examples of terms that promise uncontested precision include: 'cost–benefit', 'expected utility', 'decision theory', 'life-cycle assessment', 'ecosystem services', and 'evidence-based policy'. Yet all presuppose a set of values about what matters — sustainability for some, productivity or profitability for others[3,8]. Modellers should not hide the normative values of their choices.

Consider the value of a statistical life, loosely defined as the cost of averting a death. It is already controversial for setting

compensation — for the victims of aeroplane crashes, for instance. Although it might have a place in choosing the best public-health policy, it can produce a questionable appearance of rigour and so disguise political decisions as technical ones[8].

The best way to keep models from hiding their assumptions, including political leanings, is a set of social norms. These should cover how to produce a model, assess its uncertainty and communicate the results. International guidelines for this have been drawn up for several disciplines. They demand that processes involve stakeholders, accommodate multiple views and promote transparency, replication and analysis of sensitivity and uncertainty. Whenever a model is used for a new application with fresh stakeholders, it must be validated and verified anew.

Existing guidelines for infectious-disease modelling reflect these concerns, but have not been widely adopted[4]. Simplified, plain-language versions of the model can be crucial. When a model is no longer a black box, those using it must react to assess individual parameters and the relationships between them. This makes it possible to communicate how different framings and assumptions map into different inferences, rather than just a single, simplified interpretation from an overly complex model. Or to put it in jargon: qualitative descriptions of multiple reasonable sets of assumptions can be as important in improving insight in decision makers as the delivery of quantitative results.

Examples of models that have adhered to these guidelines can be found in forecasting flooding risk, and in the management of fisheries. These included stakeholders' insights and intuitions about both inputs and desired ends.

Mind the consequences

Quantification can backfire. Excessive regard for producing numbers can push a discipline away from being roughly right towards being precisely wrong. Undiscriminating use of statistical tests can substitute for sound judgement. By helping to make risky financial products seem safe, models contributed to derailing the global economy in 2007–08 (ref. 5).

Once a number takes centre-stage with a crisp narrative, other possible explanations and estimates can disappear from view. This might invite complacency, and the politicization of quantification, as other options are marginalized. In the case of COVID-19, issues as diverse as availability of intensive-care hospital beds, employment and civil liberties are simultaneously at play, even if they cannot be simply quantified and then plugged into the models.

Spurious precision adds to a false sense of certainty. If modellers tell the United Kingdom it will see 510,000 deaths[9] if no steps are taken to mitigate the pandemic, some might imagine a confidence of two significant digits. Instead, even the limited uncertainty analysis run by the modellers — based on just one parameter — reveals a range of 410,000–550,000 deaths. Similarly, the World Health Organization predicts up to 190,000 deaths for Africa (see go.nature.com/3hdy8kn). That number corresponds to a speculative scenario in which ten uncertain input probabilities are increased by an arbitrary 10% — as if they were truly equally uncertain — with no theoretical or empirical basis for such a choice. Although thought experiments are useful, they should not be treated as predictions.

Opacity about uncertainty damages trust. A message from the field of sociology of quantification[10] is that trust is essential for numbers to be useful[8]. Full explanations are crucial.

Mind the unknowns

Acknowledge ignorance. For most of the history of Western philosophy, self-awareness of ignorance was considered a virtue, the worthy object of intellectual pursuit — what the fifteenth-century philosopher Nicholas of Cusa called learned ignorance, or *docta ignorantia*. Even today, communicating what is not known is at least as important as communicating what is known. Yet models can hide ignorance.

"Although thought experiments are useful, they should not be treated as predictions."

Failure to acknowledge this can artificially limit the policy options and open the door to undesired surprises. Take, for instance, those that befell the heads of governments when the economists in charge admitted that their models — by design — could not predict the last recession. Worse, neglecting uncertainties could offer politicians the chance to abdicate accountability. Experts should have the courage to respond that "there is no number-answer to your question", as US government epidemiologist Anthony Fauci did when probed by a politician.

Questions not answers

Mathematical models are a great way to explore questions. They are also a dangerous way to assert answers. Asking models for certainty or consensus is more a sign of the difficulties in making controversial decisions than it is a solution, and can invite ritualistic use of quantification.

Models' assumptions and limitations must be appraised openly and honestly. Process and ethics matter as much as intellectual prowess. It follows, in our view, that good modelling cannot be done by modellers alone. It is a social activity. The French movement of *statactivistes* has shown how numbers can be fought with numbers, such as in the quantification of poverty and inequalities[11].

A form of societal activism on the relationship between models and society is offered by US-based engineer-entrepreneur Tomás Pueyo. He is not an epidemiologist, but writes about COVID-19 models and explains in plain language the implications of uncertainties for policy options.

We are calling not for an end to quantification, nor for apolitical models, but for full and frank disclosure. Following these five points will help to preserve mathematical modelling as a valuable tool. Each contributes to the overarching goal of billboarding the strengths and limits of model outputs. Ignore the five, and model predictions become Trojan horses for unstated interests and values. Model responsibly.

The authors

Andrea Saltelli is a professor at the Center for the Study of the Sciences and the Humanities, University of Bergen, Norway, and at Open Evidence Research, Open University of Catalonia, Barcelona, Spain. **Gabriele Bammer, Isabelle Bruno, Erica Charters, Monica Di Fiore, Emmanuel Didier, Wendy Nelson Espeland, John Kay, Samuele Lo Piano, Deborah Mayo, Roger Pielke Jr, Tommaso Portaluri, Theodore M. Porter, Arnald Puy, Ismael Rafols, Jerome R. Ravetz, Erik Reinert, Daniel Sarewitz, Philip B. Stark, Andrew Stirling, Jeroen van der Sluijs, Paolo Vineis.**
e-mail: andrea.saltelli@uib.no

The full list of affiliations is available online at go.nature.com/2ce8uqt. Supplementary information accompanies this Comment (go.nature.com/2ce8uqt).

1. Mayo, D. G. *Statistical Inference as Severe Testing*. (Cambridge Univ. Press, 2018).
2. Devlin, H. & Boseley, S. 'Scientists criticise UK government's 'following the science' claim' (*The Guardian*, 23 April 2020).
3. Stirling, A. 'How politics closes down uncertainty'. Available at https://go.nature.com/3kjvutz.
4. Behrend, M. R. *et al. PLoS Negl. Trop. Dis.* **14**, e0008033 (2020).
5. Kay, J. A. & King, M. A. *Radical Uncertainty: Decision-making Beyond the Numbers* (W. W. Norton & Company, 2020).
6. Puy, A., Lo Piano, S. & Saltelli, A. *Geophys. Res. Lett.* **47**, e2020GL087360 (2020).
7. Sarewitz, D., Pielke, R. A. & Byerly, R. *Prediction: Science, Decision Making, and the Future of Nature* (Island Press, 2000).
8. Porter, T. M. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton Univ. Press, 1996).
9. Ferguson, N. M. *et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand* (Imperial College London, 2020).
10. Espeland, W. N. & Stevens, M. L. *Eur. J. Sociol.* **49**, 401–436 (2008).
11. Bruno, I., Didier, E. & Vitale, T. *Partecipazione conflitto* **7**, 198–220 (2014).

BMW GROUP

Technicians process materials for battery cells at car maker BMW's research centre near Munich, Germany.

Cool metric for lithium-ion batteries could spur progress

Gregory Offer, Yatish Patel, Alastair Hales, Laura Bravo Diaz & Mohamed Marzook

A new measure for the rate of heat removal from battery packs gives manufacturers a simple way to compare products.

Lithium-ion batteries get hot, and it is hard to keep them cool. Industry has paid too little attention to this problem for the past decade. The focus has been elsewhere: on cutting costs and on boosting the amount of energy a single cell in a battery can store (energy density). This strategy has, for example, increased the longevity and capabilities of mobile phones. Future applications, such as electric vehicles and smart grids, need thousands of cells in a battery pack. These are prone to overheating.

Manufacturers of large, high-energy battery packs must design complicated systems to manage heat. The battery pack in electric-vehicle maker Tesla's Model 3 car, for example, holds more energy than 6,000 iPhone 11 handsets. Coolant fluid is pumped through a network of channels to carry heat away from the individual cells. But these cumbersome additions make the battery pack heavy and drain its energy[1]. Developers are wasting time and money on these inefficient designs. Heat-removal strategies must be improved to make battery packs both light and powerful.

Why this lack of attention? One reason is that there is no standard way of judging the thermal performance of battery packs. Manufacturers of single cells compete by chasing ever greater energy density. Their product-specification sheets do not cover how easy it is to remove heat from a cell. Designers of battery packs

thus cannot know in advance how much heat a single cell will generate. They find out too late, after investing time and money in a design.

The lithium-ion battery industry is expected to triple in size in the coming decade[2]. A step change is urgently needed in thermal management. It can be achieved quickly using proven technology.

The first step is for the battery industry to report routinely on thermal management. We have developed a standardized performance metric for this purpose[3,4]. It compares different electrochemical cells and can be measured using equipment that is readily available in battery laboratories. Including this metric on each battery specification sheet would drive competition and thus lead to improvements in single-cell designs and battery pack performance.

Thermal management

Leading car companies are investing heavily in developing better battery packs. BMW alone has put US$230 million into its battery research centre, which opened last year near Munich in Germany (see go.nature.com/2asxytj). Each company is using a different cell design and pursuing its own cooling strategy.

Broadly, there are three kinds of thermal management systems.

Air cooling. In the batteries of the Renault ZOE and Nissan LEAF car models, air is blown over the surface to remove heat. This method might be sufficient for stationary energy storage, such as for batteries that power homes, but it removes heat at a low rate. The battery packs of future electric vehicles, long-distance haulage and heavy-duty off-road vehicles will require that heat is removed faster as their performance improves year on year.

Liquid cooling. A certain volume of liquid has the capacity to remove heat about 1,000 times better than the same volume of air[5]. Cells can be immersed in flowing fluid or cooled indirectly by liquid that flows through channels wrapped around the cell. Immersion is most effective, but expensive dielectric fluids are needed to reduce the risk of a short circuit in the battery pack. Therefore, electric vehicles tend to use the cooling-channel method. Tesla wraps tubes containing liquid propylene glycol around its cylindrical cells[6]. Both immersion and cooling-channel methods drain power because of the need to pump the coolant around the battery fast enough.

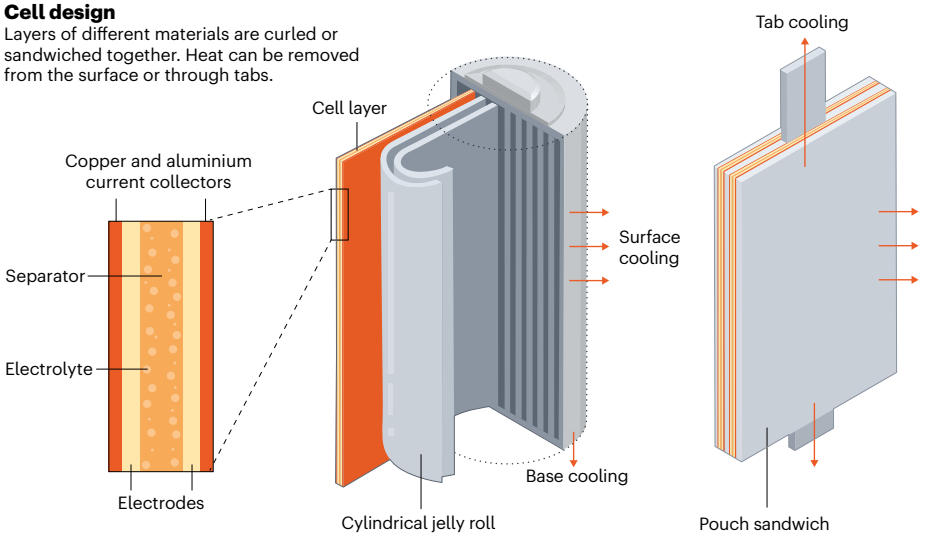
Phase-change cooling. Some materials, such as the Novec fluids made by US technology company 3M, are designed to absorb heat when they change phase — from solid to liquid or from liquid to gas — without themselves getting hotter. Cells can be immersed in or coated with such materials to absorb heat. This method is the subject of considerable research, because it uses less power and withdraws heat more evenly than do air or liquid cooling[7]. However, there is a fundamental limitation. Phase-change materials do not channel away the heat; they simply store it. Therefore, all phase-change designs require an extra cooling system to carry the heat out of the battery pack.

KEEP IT COOL

Lithium-ion batteries are prone to overheating. A metric that compares the heat removal rates of individual cells in a battery (top) could reduce the need for complicated cooling systems at the pack level (bottom).

Cell design

Layers of different materials are curled or sandwiched together. Heat can be removed from the surface or through tabs.



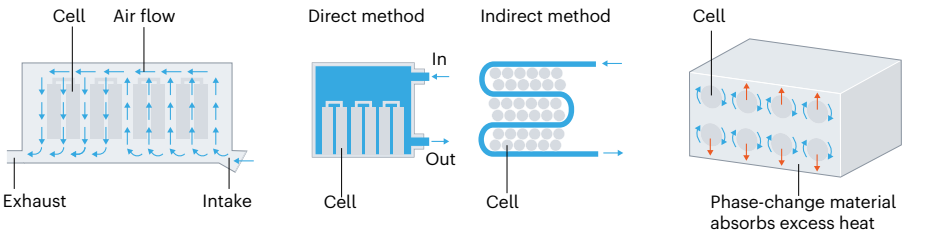
Pack management

Three different strategies can cool battery packs. Designers need to pick the best method for their application.

Air
Cheap, low power demand, poor performance

Liquid
Expensive, high power demand, good performance

Phase-change
Expensive, requires extra cooling systems, unproven performance



GRAPHIC: CLAIRE WELSH/*NATURE*

> **"Heat-removal strategies must be improved to make battery packs both light and powerful."**

Design challenge

Designers need to pick the best cooling method for their application and deploy it correctly. If they do not, the battery pack will be inefficient, supply less useful energy and degrade faster. Choosing which region of a cell to cool is the most difficult decision.

All cells are made up of layers of different materials: electrodes, an electrolyte, a separator and current collectors. The layers can be sandwiched together, as they are in pouch cells, or curled into a 'jelly roll', as in cylindrical and prismatic cells (see 'Keep it cool').

Electric current flows in and out of the cell through current collectors, which are joined to the cell's positive and negative terminals, or 'tabs'. Current collectors are made from metals which conduct heat very easily. But heat transfers slowly between the layers of the cell, because the electrodes, electrolyte and separator are thermal insulators. In other words, heat transfer parallel to the layers is faster than heat transfer across them[1].

The electrochemical performance of a cell is sensitive to temperature; at high temperatures, resistance to current flow is much lower. Thus, for the cell to be effective and stable, each layer should be exposed to identical thermal conditions. A temperature gradient between one layer and the next means that each operates slightly differently. Less energy can be taken from the cell because the hotter layer runs out of energy more quickly; some energy is left in the colder layer. The cell degrades more quickly when each layer is exposed to different rates of current flow[1,8].

Identical thermal conditions are possible only when heat is removed at the same rate from each layer. Surface cooling cannot achieve this,

because it creates a temperature gradient[1].

Removing heat through the tabs — which are connected to each layer — can cool the whole cell evenly[1,8]. Unfortunately, tab cooling is not possible in today's lithium-ion cells. Tabs are often too close to one another and too small and thin to remove enough heat from each layer. As a result, cells that are cooled through their tabs can still get dangerously hot.

Key metric

The biggest problem is more mundane. There is no thermal performance metric for electrochemical cells that is easily reproducible anywhere in the world, and that does not reveal commercially sensitive information about how a cell is designed or manufactured.

There is no good or universal method to measure cell thermal performance in the battery industry. Heat-transfer specialists favour the Biot number, which describes a body's capability to pass and dissipate heat. Mechanical engineers prefer definitions of thermal conductance and thermal conductivity; these define the rate of heat transfer that can be achieved through a material for a given temperature gradient.

None of these methods can calculate the temperature gradient across a cell when it is in operation, because electrochemical cells generate their own heat throughout their volume. If the temperature gradient across one cell is not known, it is impossible to design a thermal management system for a battery pack containing 1,000 cells.

We have developed a metric called the cell cooling coefficient[3,4]. It can be used to describe the temperature gradient across a cell in operation in watts per kelvin ($W K^{-1}$). A cell will have a different value for surface cooling and for tab cooling, because each method results in a different temperature gradient. Such a coefficient would tell a designer how difficult it will be to manage heat in the selected cells in a pack.

Our cooling coefficient is straightforward to measure in the lab. Researchers can create electrochemical heat in a cell and then determine the temperature gradient across it using temperature sensors. Heat loss from the cell can be measured using heat-flux sensors. For surface cooling, where one side of the cell is cooled and the other remains hot, the cell cooling coefficient could be calculated by dividing the rate of heat loss by the temperature gradient from the hot side to the cold side.

> **"Some will object that adding another variable will complicate protocols for optimizing cell designs."**

A large cell cooling coefficient is desirable. It means that more heat can be removed and there is a small temperature gradient inside the cell. Of the cells we have investigated, large pouch cells, such as the ones in the Nissan LEAF, seem to perform best and have a cell cooling coefficient close to 5 $W K^{-1}$ (ref. 9). Small cylindrical cells, such as the ones in the Tesla Model 3, perform less well, with a cell cooling coefficient of less than 0.5 $W K^{-1}$ (unpublished results).

Some cell manufacturers might oppose using thermal performance metrics if their products fare poorly compared to those of their competitors. Some will object that adding another variable will complicate protocols for optimizing cell designs, adding time and costs. But we estimate that this should take only an extra two hours of tests on top of the days typically spent characterizing different types of cell. And those manufacturers that embrace the metric could gain a competitive advantage.

Next steps

We call on researchers and engineers to measure and report the cell cooling coefficient routinely. Our metric should be included in publications alongside other typically reported metrics for cells, such as energy capacity and discharge rate.

Designers should evaluate thermal performance, alongside energy densities and power capabilities, to determine which cell is best suited for their battery pack. They should do this at an early stage, before designs are locked in. Computer simulations might be helpful for assessing the potential of cells. Knowing the cell cooling coefficient will help designers to evaluate trade-offs between thermal management and energy density, improving the working performance of the whole pack.

With such fierce competition in the battery industry, manufacturers who can keep their cells cool will have the brightest future.

RONNY HARTMAN/AFP/GETTY

Staff at Volkswagen in Germany assemble the lower body and battery of the ID.3 electric car.

The authors

Gregory Offer is a reader in mechanical engineering in the Electrochemical Science and Engineering Group at Imperial College London, UK, and is the principal investigator of the Faraday Institution Multi-Scale Modelling Project. **Yatish Patel** is a research fellow in mechanical engineering in the Electrochemical Science and Engineering Group at Imperial College London, UK. **Alastair Hales** is a postdoctoral researcher in mechanical engineering in the Electrochemical Science and Engineering Group at Imperial College London, UK. **Laura Bravo Diaz** is a postdoctoral researcher in mechanical engineering in the Electrochemical Science and Engineering Group at Imperial College London, UK. **Mohamed Marzook** is a PhD candidate in the Electrochemical Science and Engineering Group at Imperial College London, UK.
e-mail: a.hales@imperial.ac.uk

1. Hunt, I. A., Zhao, Y., Patel, Y. & Offer, G. J. *J. Electrochem. Soc.* **163**, A1846–A1852 (2016).
2. Choudhary, A. & Prasad, E. *Lithium-ion Battery Market by Component, End-use Industry and Automotive, and Industrial: Global Opportunity Analysis and Industry Forecast, 2019–2027* (Allied Market Research, 2020).
3. Hales, A. *et al. J. Electrochem. Soc.* **166**, A2383–A2395 (2019).
4. Hales, A., Marzook, M. W., Bravo Diaz, L., Patel, Y. & Offer, G. *J. Electrochem. Soc.* **167**, 020524 (2020).
5. Rogers, G. & Mayhew, Y. *Engineering Thermodynamics: Work and Heat Transfer* (Longman Scientific & Technical, 1992).
6. Adams, D. T. *et al.* Battery Pack Thermal Management System. US patent 20090023056A1 (2009).
7. Ianniciello, L., Biwolé, P. H. & Achard, P. *J. Power Sources* **378**, 383–403 (2018).
8. Zhao, Y., Patel, Y., Zhang, T. & Offer, G. J. *J. Electrochem. Soc.* **165**, A3169–A3178 (2018).
9. Dondelewski, O. *et al. eTransportation* (in the press).

Correspondence

Europe: yes to philanthropy

The COVID-19 pandemic has highlighted the limited coordination among European countries when they need to act quickly and forcefully. As scientists and national academy members from 15 European countries, we endorse a call for a European Foundation for the Prevention of Environmental and Health Crises. Its aim would be to forge an alliance between European scientists and philanthropists so that we are better prepared for our shared future.

Individual donors and foundations established by benefactors account for 1.95% of gross domestic product (GDP) in the United States, but only 0.65% of the GDP of the European countries for which data are available (see go.nature.com/2ykyarb).

The new European body would be similar to the Bill & Melinda Gates Foundation set up in Seattle, Washington, 20 years ago. It would need an initial endowment of about €20 billion (US$22.5 billion) from European philanthropists (see https://europe-foundation.eu).

With advice from an independent scientific advisory committee, the foundation's council of donors would support European research into technology and therapeutics to help counter epidemics and environmental threats in the future.

Francis-André Wollman* Institute of Physico-Chemical Biology, Paris, France.
wollman@ibpc.fr
*On behalf of 75 co-signatories; for a full list see go.nature.com/2uxaydx.

Space leaders urge COVID-19 unity

As a former head of China's National Space Science Center and former directors of the European Space Agency (ESA) and NASA, we call for a one-planet approach to tackling the COVID-19 pandemic.

Coming from three continents, we are united by our passions for space — as an engineer, a scientist and an astronaut. We are crew members of Planet Earth and friends for life. As such, we appreciate the power of a holistic vision of our world.

Since astronauts first shared their space-based views of Earth and its thin, fragile atmosphere, we have felt a responsibility to protect the planet. Many of those pioneer astronauts became environmentalists, advocates for international cooperation and sustainable development, and world-peace envoys. As space exploration moved into the Apollo era and beyond, we discovered that thousands of exoplanet systems could exist. Some might even host intelligent life. More people will come to share these new world views as space tourism takes off.

We call for a global effort to end the pandemic and to then shape the world into a better place for international cooperation and sustainable development.

Ji Wu National Space Science Center, Chinese Academy of Sciences, Beijing, China.
wuji@nssc.ac.cn

Jean-Jacques Dordain European Space Agency, Paris, France.

Charles Bolden NASA, Washington DC, USA.

Racism: words are fine, now act

You will appreciate the scepticism in Black and minority-ethnic researchers (whose experiences I study) about whether statements such as yours on ending systemic racism in science (see *Nature* **582**, 147; 2020) will be followed through, or just dissolve as the news cycle moves on — as tends to happen.

If *Nature* is sincere, it must do four things now.

The first is to detail how systematic racism in science has operated throughout history. This can no longer be denied — from the erasure of the scientific achievements of scholars who are not white or from Western countries, to the enduring application of racist classification systems to people outside Europe, their languages and ideas.

The second is to admit the complicity of the institutions of science in the oppression and subjugation of populations through colonial and imperialist actions. These were led by (but were not exclusive to) the state and corporations of Britain such as the British East India Company, and their equivalents in other European nations.

Steps three and four are follow-up actions. Issue regular reviews of the diversity of staff, as well as of the scientists who submit to, review and publish in the journal. And scrutinize the scope of the publication: it should be truly reflective of its global reach.

So, what will *Nature* as an institution now do to address the injustices that you have highlighted so eloquently?

Fauzia Ahmad Department of Sociology, Goldsmiths, University of London, UK.
f.ahmad@gold.ac.uk

From 1950s malaria to COVID-19

News coverage of the COVID-19 epidemic makes frequent reference to the reproduction number, R_0, the average number of new cases of a disease that arise from a single case. As well as recognizing its simple mathematical power and the challenges its use poses (see C. Uzoigwe *Nature* **582**, 341; 2020), it is important to understand how it originated.

The R_0 concept has been attributed to the late Robert May (1936–2020). Although May championed R_0 and contributed to its application (see R. M. Anderson and R. M. May (eds) *Population Biology of Infectious Diseases*; Springer, 1982), it was first developed more than 60 years ago by the epidemiologist George Macdonald, then director of the Ross Institute of the London School of Hygiene and Tropical Medicine. His aim was to understand quantitatively the transmission of malaria, a mosquito-borne disease (G. Macdonald *The Epidemiology and Control of Malaria*; Oxford Univ. Press, 1957). He derived R_0, originally designated Z_0, from a reproduction ratio established by the demographer Alfred J. Lotka (see D. L. Smith *et al. PLoS Pathog.* **8**, e1002588; 2012) .

The number became known as R_0 in the 1970s, and has since been widely applied in disease epidemiology.

Jake Baum, **Geoffrey Pasvol** Imperial College London, UK.
jake.baum@imperial.ac.uk

Richard Carter University of Edinburgh, UK.

News & views

Parkinson's disease

Unleashing the neuronal side of astrocyte cells

Ernest Arenas

Astrocytes are non-neuronal brain cells that express a protein called PTB. It emerges that PTB depletion unlocks the potential of astrocytes to convert to neurons in a mouse model of Parkinson's disease. See p.550

How can the neurons that degenerate in the substantia nigra region of the midbrain in Parkinson's disease be replaced? One exciting possibility would be to convert non-neuronal cells called astrocytes, which are plentiful in the brain, into neurons. On page 550, Qian *et al.*[1] report a simple strategy that harnesses this possibility and can ameliorate neurological deficits in a mouse model of Parkinson's disease[1]. Their work, along with a parallel approach recently outlined by Zhou *et al.* in *Cell*[2], holds huge promise for our ability to use cell-conversion strategies to treat neurodegenerative diseases.

Cell types such as skin cells or astrocytes can be converted – through forced expression of transcription factors, microRNAs or small molecules – to other cell types *in vitro*[3–8], including to neurons that produce the neurotransmitter molecule dopamine[5,8]; these neurons are lost in Parkinson's disease. This approach has also been used to convert mouse-brain astrocytes to neurons *in vivo*[6–8]. For instance, astrocytes in the brain's striatum have been converted to 'induced dopamine-releasing' (iDA) neurons that can partially correct motor defects in a mouse model of Parkinson's disease[8]. However, the iDA neurons generated using this approach neither formed the distant neuronal connections found in a healthy brain nor comprehensively restored motor behaviour. Qian *et al.* and Zhou *et al.* have used an alternative strategy to efficiently reprogram astrocytes into neurons: depletion of an RNA-binding protein called PTB that is expressed in astrocytes and that inhibits neuronal differentiation.

Qian *et al.* began their experiments *in vitro*, using astrocytes isolated from the cortex and midbrain of mouse brains, and from the human cortex. The authors used an RNA molecule called a small hairpin RNA to promote degradation of messenger RNA transcribed from the gene that encodes PTB, *Ptbp1*. This triggered conversion of all three types of astrocyte to neurons. Zhou *et al.* achieved the same effect using the genome-editing technique CRISPR–CasRx to deplete *Ptbp1* mRNA in astrocytes isolated from the mouse cortex.

Next, the two teams depleted PTB *in vivo* in the adult mouse brain. Qian and colleagues used mice genetically engineered such that astrocytes could be targeted by the small hairpin RNA against *Ptbp1*, which was carried into the brain in a viral construct. By contrast, Zhou *et al.* infected astrocytes in wild-type mice with a virus that carried the CRISPR–CasRx machinery. Both strategies led to conversion of the targeted astrocytes to neuronal cell types.

The groups next depleted PTB in a mouse model of Parkinson's disease. In these animals – as in people who have the disorder – dopamine-releasing neurons are depleted in the substantia nigra, and dopamine levels are abnormally low in the striatum (the area to which these neurons project), resulting in deficits in motor behaviour. Qian *et al.* depleted PTB in astrocytes in the substantia nigra of these animals; Zhou *et al.* in the striatum. Both approaches yielded the same result: conversion of some of the infected astrocytes to neurons that resembled those lost in Parkinson's disease, and restoration of motor behaviour.

The two groups demonstrated that PTB depletion causes astrocytes to convert to neuronal cell types largely appropriate to the brain region they reside in. How is this specificity conveyed? Qian *et al.* found that astrocytes in the midbrain express low levels of the transcription

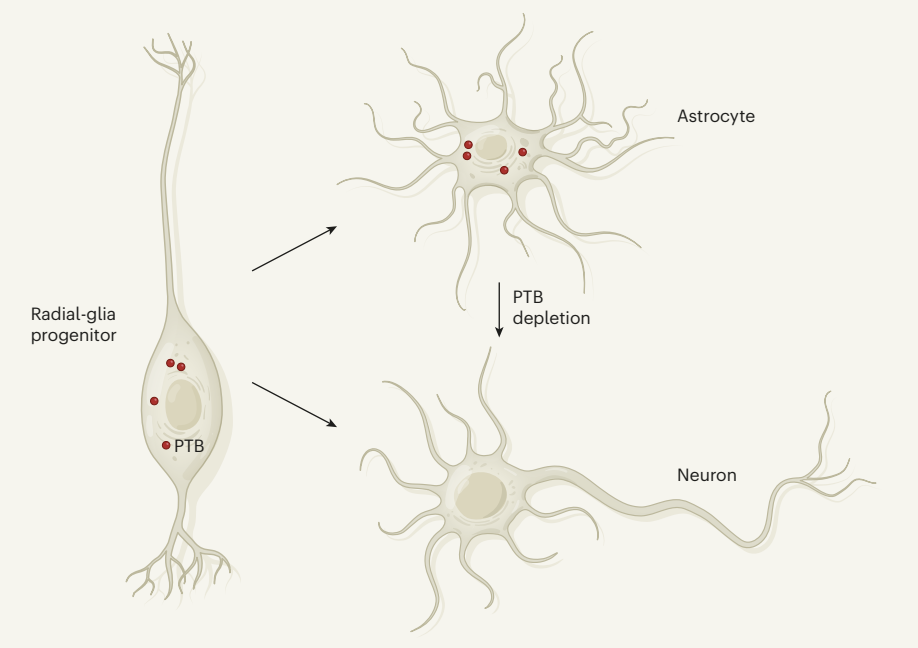


Figure 1 | A path for astrocyte-to-neuron conversion. Radial-glia progenitors are stem-cell-like cells that – in the developing mouse midbrain – express messenger RNA that encodes the protein PTB. These cells can give rise to non-neuronal cells called astrocytes, which also express this mRNA, and to neurons, which do not. Two groups[1,2] report that depletion of PTB in adult astrocytes leads to the cells' conversion to neurons. In a mouse model of Parkinson's disease, PTB depletion produced the type of dopamine-releasing neuron that is lost in Parkinson's disease, and restored motor behaviour (not shown).

factors Lmx1a and Foxa2; these are expressed in the progenitors of dopamine-releasing neurons during midbrain development and are required for the maturation of these progenitors into neurons[9]. PTB depletion further increased the expression of these factors in midbrain astrocytes. By contrast, in cortical astrocytes, the treatment led to increased levels of transcription factors associated with cortical neurons, such as Ctip2 and Cux. In addition, reprogramming of astrocytes in the substantia nigra, or in the neighbouring ventral tegmental area, produced different subtypes of iDA neuron that express subtype-specific transcription factors and proteins: Sox6 and Aldh1a1 in the substantia nigra, Otx2 in the ventral tegmental area.

Qian and colleagues' results indicate that brain-region-specific transcription factors contribute to the astrocyte-to-iDA conversion. However, such a mechanism cannot explain why Zhou *et al.* were able to convert striatal astrocytes to iDA neurons, given that striatal astrocytes express a different set of region-specific transcription factors. What might be the mechanism leading to iDA conversion in the striatum?

Zhou and colleagues show an almost threefold increase in iDA conversion efficiency in the mouse model of Parkinson's disease compared with control mice one month after treatment. These results suggest that astrocytes themselves, or cells in their environment, respond to the loss of endogenous dopamine-releasing neurons by expressing factors that promote the conversion of astrocytes to iDA neurons. And Qian *et al.* found higher conversion efficiency in the mouse midbrain than in isolated midbrain astrocytes, indicating a role for local brain-derived factors in iDA conversion. Identifying local and damage- or disease-specific factors, intrinsic or extrinsic to cells, holds the key to further improving the efficiency of astrocyte-to-neuron conversion.

One intriguing question to arise from these studies is why astrocytes are constantly repressing neuronal genes. One explanation might lie in the cells' developmental origin. Astrocytes and neurons have common ancestors called radial-glia progenitors — stem-cell-like cells that first give rise to neurons and then differentiate into astrocytes and other neuron-supporting glial cells[10]. In the developing mouse midbrain, all radial-glia cell types express *Ptbp1*, whereas differentiating neuron precursors and neurons do not[11]. Perhaps midbrain astrocytes — as descendants of radial glia — have inherited a program to generate neurons that lies dormant unless PTB is depleted (Fig. 1). *Ptbp1* is also expressed in other midbrain cell types[11], including endothelial and pericyte cells in the blood vessels, ependymal cells lining the ventricular cavity and immune cells called microglia. Future studies should examine whether PTB depletion can also convert these cells to iDA neurons in animal models of Parkinson's disease.

For this strategy to be useful in the clinic, its efficiency might need to be improved. For instance, 60–65% of the infected astrocytes do not become iDA neurons. This percentage must decrease, either through more-focused targeting of astrocytes in the substantia nigra, or by introducing factors that enable non-nigral astrocytes to convert to iDA neurons. It will also be important to determine the quality and authenticity of the converted iDA cells at single-cell level, and to investigate whether unwanted cells are generated. Both Qian *et al.* and Zhou *et al.* provide evidence that astrocytes are converted to other neuron types, besides iDA cells. Moreover, Qian *et al.* show that converted iDA neurons mainly project to the septum, rather than the striatum, and that only 8% of the fibres that project to the septum come from iDA neurons. However, on a positive note, more than half of the fibres reaching the striatum were contributed by iDA neurons. This finding — together with the demonstration that the conversion process restored striatal dopamine levels and motor activity — provides evidence for a remarkable functional reconstitution of the nigrostriatal pathway by iDA neurons.

In a final set of experiments, Qian *et al.* explore a way in which their approach might be used in the clinic: using short nucleic acids called antisense oligonucleotides that bind to an mRNA and prevent its translation into protein. The authors show that local transient delivery of antisense oligonucleotides against PTB led to the generation of iDA-like neurons and to motor recovery in the mouse model of Parkinson's disease, demonstrating the validity of the approach.

Future experiments will need to examine whether human midbrain or striatal astrocytes can also be converted to iDAs, and whether the converted cell types and their targets are correct and stable over long periods. The safety of PTB depletion and the strategies used to deliver the treatment will also have to be carefully assessed, to rule out any collateral damage to bystander host brain cells or to the converted cells, or any damage resulting from the strategy's depletion of astrocytes. Although many questions remain to be answered, the simplicity and efficiency of this gene-therapy approach to cell replacement makes it very attractive. The current studies promise to open a new chapter in the development of regenerative medicine for neurological disorders such as Parkinson's disease.

Ernest Arenas is in the Division of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institute, Stockholm 17177, Sweden.
e-mail: ernest.arenas@ki.se

1. Qian, H. *et al. Nature* **582**, 550–556 (2020).
2. Zhou, H. *et al. Cell* **181**, 590–603.e16 (2020).
3. Takahashi, K. & Yamanaka, S. *Cell* **126**, 663–676 (2006).
4. Vierbuchen, T. *et al. Nature* **463**, 1035–1041 (2010).
5. Caiazzo, M. *et al. Nature* **476**, 224–227 (2011).
6. Heins, N. *et al. Nature Neurosci.* **5**, 308–315 (2002).
7. Gascón, S., Masserdotti, G., Russo, G. L. & Götz, M. *Cell Stem Cell* **21**, 18–34 (2017).
8. Rivetti di Val Cervo, P. *et al. Nature Biotechnol.* **35**, 444–452 (2017).
9. Arenas, E., Denham, M. & Villaescusa, J. C. *Development* **142**, 1918–1936 (2015).
10. Kriegstein, A. & Alvarez-Buylla, A. *Annu. Rev. Neurosci.* **32**, 149–184 (2009).
11. La Manno, G. *et al. Cell* **167**, 566–580 (2016).

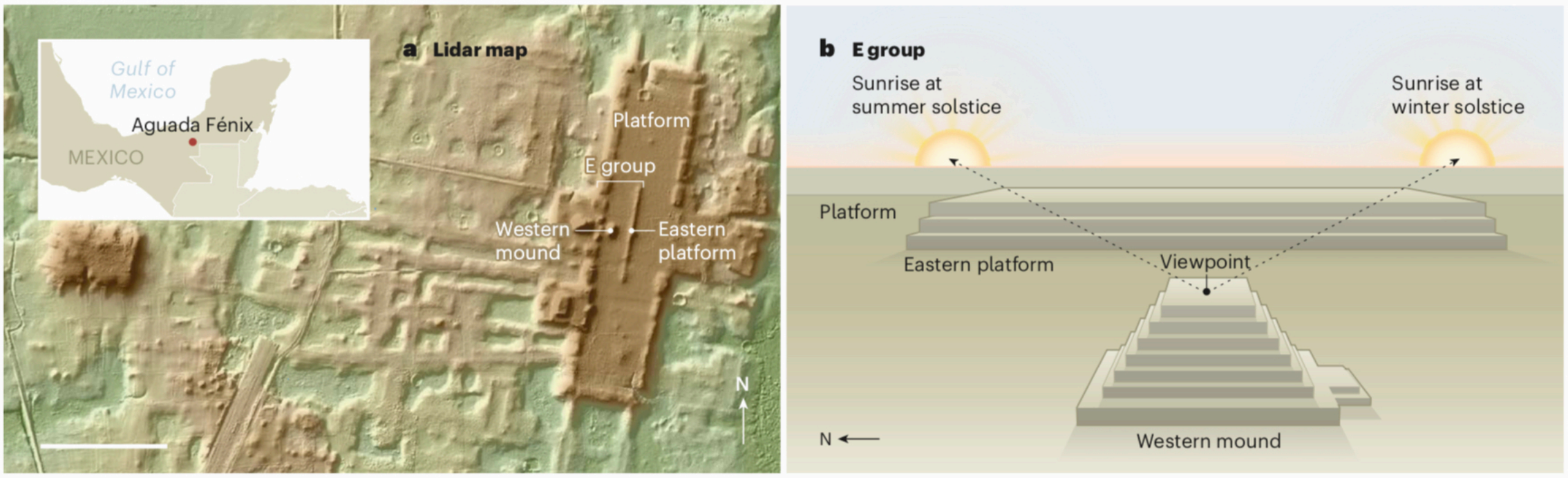
Archaeology

Large-scale early Maya sites revealed by lidar

Patricia A. McAnany

Archaeology is transforming our view of how ancient Maya societies developed. Use of lidar technology has now led to the discovery that large, monumental structures that aid naked-eye astronomy were built unexpectedly early. **See p.530**

In archaeology, there are few watershed moments, when a technological breakthrough changes everything. But the invention of radiocarbon dating in the 1940s brought one such revolution, by providing a consistent, worldwide system for placing archaeological material in chronological order. A more-recent transformative innovation is the airborne application of a remote-sensing technique called light detection and ranging (lidar) to create a model (also known as a digital-elevation model) of the bare-surface terrain that is hidden by trees in forested areas[1]. Lidar is changing archaeological study of the ancient Maya in Mexico and Central America. It is increasing the speed and scale



A. IMAGE, TAKESHI INOMATA

Figure 1 | An early Maya site. a, Inomata *et al.*[2] report the discovery of a site in Mexico at Aguada Fénix that is associated with the ancient Maya. Using surface-mapping technology called lidar, followed by excavations, the authors reveal a huge platform built from clay and earth that dates to 1000–800 BC. It contains a type of structure called an E Group (comprising a western mound and eastern platform) that is associated with astronomical observations. Inomata and colleagues' finding reveals surprisingly early large-scale landscape alterations pre-dating the emergence of Maya royal courts and providing insight into how Maya societies developed. Scale bar, 500 metres. **b**, In a typical E-Group arrangement, a western mound or pyramid provides a viewing site that aids the observation of sunrise on the horizon at the summer and winter solstices. These events are viewed by looking towards the corners of an elongated platform to the east.

of discovery, and reshaping our understanding of the antiquity of monumental-scale landscape alteration. On page 530, Inomata *et al.*[2] provide a prime example of this in their study of the region of Tabasco in Mexico that borders the Usumacinta River.

Lidar requires an aeroplane or drone to fly over the area of interest. Laser pulses are emitted and signals bouncing back generate what is termed a point cloud of data points. Expert image processing and prodigious computer capacity can then yield models of bare terrain from which the vegetation has been digitally removed. In areas where dwellings, platforms, pyramids and even palaces can be obscured by high-canopy vegetation, a bare-terrain model yields something close to a topographic map of the surface. Straight lines and corners in a bare-terrain model suggest elements that have human rather than geological origins.

Generating such models might not sound impressive for arid landscapes, but it is a game changer where high-canopy trees obscure the view. Lidar images from one plane flight can provide more information than can be generated by decades of conventional archaeological surveys. As a veteran of pre-lidar survey techniques and an archaeologist who works in the humid tropics that are associated with ancient Maya civilizations, I have spent thousands of hours of fieldwork walking behind a local machete-wielding man who would cut straight lines through the forest. This process creates a grid within which we archaeologists proceed on foot to locate any structures present. Then, after more machete-cutting to reveal the corners, shape and height of ancient constructions, the structures could finally be mapped.

This time-consuming process has required years, often decades, of fieldwork to map a large ancient Maya city such as Tikal in Guatemala and Caracol in Belize. At Caracol, laborious clearing and mapping were under way for decades before lidar quickly revealed the full extent of the agricultural terraces and settlement[1]. Bare-terrain models produced by lidar imagery include coordinate information (such as latitude and longitude) that can be used to 'ground-truth' the results by examining the specific physical site. Machete-cutters are still needed during this ground-truthing step.

Airborne lidar has been of benefit for the study of other archaeological sites in tropical forests, such as those at Angkor Wat in Cambodia[3]. Lidar data have revealed artificial reservoirs built around the temples there, yielding subtle hints about the limits to the resilience of this complex hydraulic system. This research has also underlined the vastness of the landscape modifications undertaken by people of the Khmer Empire[4].

Back in the tropical forests of Petén in Guatemala — the heartland of Maya 'divine' rulers during the Classic period (AD 250–800) — is a region called the Maya lowlands that archaeologists have studied nearly continuously since the mid-twentieth century. Intriguing because of its hieroglyphic writing system, naturalistic sculpture and painting style, and adroitness in maize (corn) farming, Classic-period Maya society was organized politically around dozens of royal courts. Archaeologists have lavished much attention on these courts, but only a small fraction of the landscape beyond and between them has been mapped using conventional methods.

To remedy this, a large lidar programme was initiated[5]. The resultant bare-terrain models show a landscape that was intensively and deliberately modified by humans in a way that would easily have escaped detection by even the most seasoned conventional field mappers. The spatial continuity of landscape modifications can be more obvious when viewed from above rather than at ground level. Inomata and colleagues echo this point in reference to their key finding: the discovery of massive, ancient platforms made of clay and earth, measuring about 400 metres across and 1,400 metres in length, at Aguada Fénix in the Usumacinta region of Tabasco, which lies at the western boundary of the Maya lowlands. These platforms date to between 1000 and 800 BC. In the northern part of the Maya lowlands, where the forest is more scrubby but no less impenetrable than the Tabasco forest, a similar lidar 'reveal' is happening with equally profound results[6–9].

In comparison with the Maya region farther east and the region to the west associated with Olmec societies (known for colossal stone heads from the second to first millennium BC), Tabasco has taken a back seat in terms of archaeological investigations during the past century, despite its position between those two regions. That changed when Inomata and colleagues decided to conduct a survey in twenty-first-century style using lidar. This was not a random 'fishing expedition' to discover whatever they could find. Instead, their search focused on a type of construction called an E Group (Fig. 1). Known as the earliest form of non-residential architecture in the Maya lowlands, E Groups were used for naked-eye astronomy[10]. Some, such as those found by Inomata and colleagues, were built up to 3,000 years ago and, interestingly, they pre-date even a clear footprint of settlement in the form of dwellings and villages.

Archaeologists seek to understand which came first in the development of community life — sedentary life in a fixed dwelling, or periodic gatherings for group-based ritual activities, such as religious or astronomical observances. The former was generally

thought to have paved the way for the latter, but newer evidence is emerging to suggest it was the other way around.

Human ancestors might first have come together to mark the change of seasons observable in the movement of the Sun or other celestial bodies across the sky or along the horizon. E Groups (Fig. 1) contain a low mound or pyramid on the western side of an architectural complex with an elongated platform on the eastern side. Looking from the western structure aids the viewer to witness sunrise during the winter and summer solstices, which are visible along the northern and southern corners, respectively, of the eastern platform (which is elongated from north to south). Brilliantly simple in design, this type of construction was built, over and over again, up and down the Usumacinta region and throughout the Maya lowlands to the east.

Using the revealing 'eyes' of lidar, Inomata and colleagues document 16 instances of E-Group constructions during the first millennium BC. These were built on top of massive rectangular platforms. The platform at Aguada Fénix is the largest of any such platform discovered from this early time period, and Inomata and colleagues suggest that it might be the largest Maya construction built before Spanish invaders arrived. On the basis of the site's absence of excavated stone sculpture depicting rulers — such as the colossal heads found from the same time period in the Olmec region — the authors argue that these constructions were truly public architecture and not built at the behest of rulers. If so, then why were they built so large, and abandoned only hundreds of years later (as indicated by radiocarbon-dating information from the authors' excavations)? And how far to the east and west of Aguada Fénix can such arrangements of a huge platform with an E Group be found? Strictly speaking, this architectural pattern is not a strong characteristic of the central Maya lowlands to the east nor of the Olmec region to the west.

Many questions remain for further research, but there is no doubt that lidar is continuing to transform archaeological research in forested regions. At Aguada Fénix, in particular, the lidar data coupled with Inomata and colleagues' excavations substantially deepen our understanding of the social transformations that occurred there, and strengthen the argument that public architecture on a monumental scale pre-dated village life in eastern Mesoamerica. These findings will lead some to cast a critical eye on the proposed link between public architecture and hierarchical rulership, given that the latter seems to have commenced in the Maya lowlands hundreds of years after the construction of the Aguada Fénix site. The fact that Inomata and colleagues' research took three years, rather than three decades, also demonstrates the powerful way in which lidar is facilitating the rapid detection and investigation of the past by offering a way of peering through the veils of the forest canopy.

Patricia A. McAnany is in the Department of Anthropology, University of North Carolina, Chapel Hill, North Carolina 27599, USA.
e-mail: mcanany@email.unc.edu

1. Chase, A. F., Chase, D. Z., Fisher, C. T., Leisz, S. J. & Weishampel, J. F. *Proc. Natl Acad. Sci. USA* **109**, 12916–12921 (2012).
2. Inomata, T. *et al. Nature* **582**, 530–533 (2020).
3. Evans, D. H. *et al. Proc. Natl Acad. Sci. USA* **110**, 12595–12600 (2013).
4. Evans, D. *J. Archaeol. Sci.* **74**, 164–175 (2016).
5. Canuto, M. A. *et al. Science* **361**, eaau0137 (2018).
6. Hare, T., Masson, M. & Russell, B. *Remote Sensing* **6**, 9064–9085 (2014).
7. Hutson, S. R., Kidder, B., Lamb, C., Vallejo-Cáliz, D. & Welch, J. *Adv. Archaeol. Pract.* **4**, 268–283 (2016).
8. Magnoni, A. *et al. Adv. Archaeol. Pract.* **4**, 232–248 (2016).
9. Stanton, T. W. *et al. J. Archaeol. Sci. Rep.* **29**, 102178 (2020).
10. Friedel, D. A., Chase, A. F., Dowd, A. S. & Murdock, J. (eds) *Maya E Groups: Calendars, Astronomy, and Urbanism in the Early Lowlands* (Univ. Press Florida, 2017).

This article was published online on 3 June 2020.

Cancer

Tumour metabolites hinder DNA repair

Lei-Lei Chen & Yue Xiong

Altered metabolism and genome instability are hallmarks of cancer. A mechanism now explains how three small molecules that accumulate in tumours connect abnormal metabolism to genomic problems by hindering DNA repair. **See p.586**

Towards the end of the nineteenth century, chromosomal abnormalities detected under the light microscope revealed that a type of massive genome instability resulting in an abnormal number of chromosomes occurs in certain types of cancer. Not long after, the biochemist Otto Warburg observed that tumour cells tend to use pathways of glucose and energy metabolism that are distinct from those used by normal cells. We now know that genome instability and altered metabolism are two common characteristics of most tumour cells. Genome instability has been investigated continuously since its discovery; altered metabolism was rediscovered as a research area only recently. But not much crosstalk between these two processes in cancer has been reported so far. Sulkowski *et al.*[1] reveal on page 586 how several metabolites that accumulate to high levels in tumour cells suppress DNA repair, thus revealing a direct link between altered metabolism and genome instability caused by DNA damage.

Mutations targeting the genes encoding the enzymes isocitrate dehydrogenase 1 and 2 (IDH1 and IDH2) result in cells accumulating high levels of the metabolite 2-hydroxyglutarate (2-HG). Mutations in the genes encoding the enzymes fumarate hydratase and succinate dehydrogenase cause cells to accumulate high levels of the molecules fumarate and succinate, respectively. These three small molecules are often referred to as oncometabolites because their accumulation boosts tumour development[2,3], and they are structurally similar to the molecule α-ketoglutarate (α-KG). This is an intermediate in the Krebs-cycle pathway that also serves as a component, called a co-substrate, needed for the function of a family of enzymes called α-KG/Fe(II)-dependent dioxygenases.

This enzyme family, which comprises 65 members in humans[4], catalyses a diverse range of oxidation reactions in proteins, DNA, RNA and lipids. In these reactions, α-KG binds to the active site of the enzyme to aid catalysis. However, 2-HG, succinate and fumarate can compete with α-KG for binding to this catalytic site and thus inhibit these enzymes. One such enzyme is lysine histone demethylase (KDM), which modifies chromatin — the complex of DNA and proteins of which chromosomes are made[5–7].

Two closely related KDMs, called KDM4A and KDM4B, catalyse the removal of a methyl group (demethylation) from a lysine amino-acid residue (termed K9) in the DNA-binding histone 3 (H3) proteins in chromatin. The methylation of H3K9 is linked to a pathway called the homology-dependent repair (HDR) pathway, which mends double-strand breaks (DSBs) in DNA[8]. DSBs are the most dangerous type of DNA damage. If left unrepaired, they can cause chromosome breakage and genomic instability that might promote tumour growth or lead to cell death.

Sulkowski and colleagues investigated HDR in human cancer cells grown *in vitro*. They found that, at a DSB site, the local addition of three methyl groups to H3K9 to generate

trimethylated H3K9me3 residues has a key role in the initiation of HDR. In tumour cells that have mutations in the genes encoding IDH1, IDH2, fumarate hydratase or succinate dehydrogenase, the authors report that high levels of oncometabolites inhibit KDM4B. This inhibition of demethylation results in a widespread hypermethylation of H3K9 that masks the specific local appearance of H3K9me3 marks and impairs the recruitment of factors needed for HDR and DSB repair (Fig. 1).

A link between oncometabolites and DNA-repair defects was previously suggested by the clinical finding that people who have a type of cancer called glioma with mutations in the *IDH1* or *IDH2* genes benefited from a combination of chemotherapy and radiation therapy, both of which induce DNA damage[9]. That finding indicates that tumours that accumulate high levels of oncometabolites are vulnerable to therapy that causes DNA damage. Moreover, a genomic analysis of different types of cancer ranked *IDH1* as being the fifth most frequently mutated human gene that is connected to DNA repair[10].

Two mechanisms have previously been proposed to explain how the 2-HG that accumulates when *IDH1* or *IDH2* are mutated causes DNA-repair defects. One idea is that 2-HG directly inhibits the enzymes ALKBH2 and ALKBH3, which repair methylation-induced single-strand DNA damage[11]. Another suggestion is that 2-HG inhibits H3K9 demethylases and thereby causes a reduction in the expression of ATM, a key protein required for DNA repair[12].

Sulkowski and colleagues had previously found that oncometabolites suppressed the HDR pathway and had identified KDM4A and KDM4B as being important for DSB repair[11]. The authors therefore explored possible connections between these processes. HDR is a complex event that involves the sequential recruitment of multiple repair factors to DSB sites, with the protein Tip60 being among the first to arrive at the damaged region[8]. Sulkowski *et al.* used a system in which human cells grown *in vitro* were engineered to allow the precise initiation of DSB and monitoring of the repair process.

The authors found that in control cells that did not have high levels of oncometabolites, a rapid spike of H3K9me3 modifications occurred locally in chromatin in the vicinity of the DSB within 30 minutes of the DSB being induced. This was followed by the coordinated recruitment of factors needed for HDR. However, in cancer cells with high levels of oncometabolites, H3K9me3 was elevated throughout the genome before DSB induction, and the subsequent recruitment of the factors needed for HDR was substantially impaired compared with that in the control cells. These defects in repair-factor recruitment could be prevented by deleting the mutant version of *IDH1* or by treatment with a pharmacological inhibitor of mutant IDH1 protein to block 2-HG production. These results establish a causal relationship between the presence of oncometabolites and impaired DSB repair.

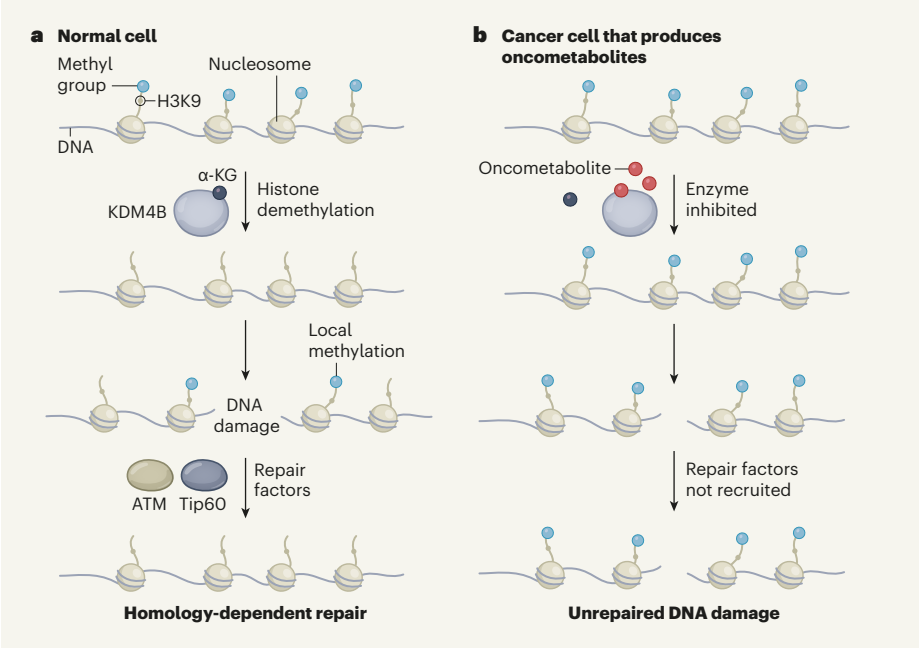


Figure 1 | How molecules in cancer cells inhibit the repair of DNA damage. a, DNA wraps around histone proteins to form a structure called a nucleosome. In normal cells, the enzyme KDM4B catalyses the removal of methyl groups from the lysine 9 (K9) amino-acid residue of the protein histone 3 (H3) in the nucleosome. This H3K9 demethylation activity requires the small molecule α-ketoglutarate (α-KG). If a double-strand break in DNA occurs, H3K9 is methylated at the damage site and this local methylation signal recruits DNA-repair factors that include the proteins Tip60 and ATM. These fix the damage through a process called homology-dependent repair. **b**, As a result of certain mutations, some cancer cells accumulate small molecules termed oncometabolites that promote tumour growth. Sulkowski *et al.*[1] have revealed a mechanism that underlies this phenomenon. Oncometabolites compete with α-KG for binding to KDM4B and thus inhibit the enzyme's function. This results in H3K9 methylation across the genome. This global hypermethylation masks a local spike in H3K9 methylation occurring after DNA damage, and hinders the recruitment of DNA-repair factors. Unrepaired DNA damage can lead to genome instability and thus boost tumour growth.

How might KDM4B inhibition by oncometabolites impair HDR? Local H3K9 methylation activates Tip60, which in turn activates ATM, a key enzyme needed for HDR. Results from a series of experiments support the authors' model that a sudden increase in H3K9me3 modifications at a DSB site serves as a key signal to recruit repair factors. Blocking the accumulation of oncometabolites, adding α-KG, or engineering cells to express KDM4A or KDM4B (but not other KDMs or ALKBH2 or ALKBH3), resulted in a decrease in global genomic H3K9me3 modifications and restored both the recruitment of repair factors and DSB repair at an engineered DNA-damage site, compared with the effects seen in cells that did not receive such treatment. If cells producing oncometabolites were engineered to have a mutant version of a histone that sequesters H3K9 methyltransferase enzymes and thus reduces the genomic level of H3K9me3 modifications, the cells displayed an H3K9me3 spike on DSB formation that led to Tip60 recruitment and repair of DNA damage.

Sulkowski and colleagues' findings expand the known roles of oncometabolites and raise several interesting questions. How does the rapid spike in H3K9me3 at a DSB site result in the coordinated recruitment of repair proteins, and what factor(s) might recognize such a modification of a DSB site? Around the DSB site, does hypermethylation of H3K9, which is known to recruit repressive factors that drive the formation of a condensed form of chromatin called heterochromatin, prevent the binding of factors needed for HDR? Questions also remain about whether the roles of KDM4A and KDM4B differ in HDR. Both enzymes catalyse the same type of H3K9 demethylation, and boosting their expression can overcome inhibition by oncometabolites and prevent HDR defects. Yet the authors report that the depletion only of KDM4B impairs HDR.

The enzyme PARP promotes the repair of single-strand DNA breaks, and inhibitors that block PARP are used to treat certain types of cancer. Tumour cells that produce 2-HG are particularly prone to death if treated with PARP inhibitors[11]. The findings by Sulkowski *et al.* might lead to new therapeutic strategies that exploit the therapeutic opportunities

arising from oncometabolite accumulation, given that we now have a clearer picture of how such cancer cells are vulnerable if DNA-repair processes are targeted.

Lei-Lei Chen and **Yue Xiong** are in the Department of Biochemistry and Biophysics, and at the Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, North Carolina 27516, USA.
e-mail: yxiong@email.unc.edu

1. Sulkowski, P. L. *et al. Nature* **582**, 586–591 (2020).
2. King, A., Selak, M. A. & Gottlieb, E. *Oncogene* **25**, 4675–4682 (2006).
3. Ye, D., Guan, K.-L. & Xiong, Y. *Trends Cancer* **4**, 151–165 (2018).
4. Rose, N. R., McDonough, M. A., King, O. N. F., Kawamura, A. & Schofield, C. J. *Chem. Soc. Rev.* **40**, 4364–4397 (2011).
5. Chowdhury, R. *et al. EMBO Rep.* **12**, 463–469 (2011).
6. Xu, W. *et al. Cancer Cell* **19**, 17–30 (2011).
7. Xiao, M. *et al. Genes Dev.* **26**, 1326–1338 (2012).
8. Sun, Y. *et al. Nature Cell Biol.* **11**, 1376–1382 (2009).
9. Cairncross, J. G. *et al. J. Clin. Oncol.* **32**, 783–790 (2014).
10. Knijnenburg, T. A. *et al. Cell Rep.* **23**, 239–254.e6 (2018).
11. Sulkowski, P. L. *et al. Sci. Transl. Med.* **9**, eaal2463 (2017).
12. Inoue, S. *et al. Cancer Cell* **30**, 337–348 (2016).

This article was published online on 3 June 2020.

Quantum physics

A step closer to secure global communication

Eleni Diamanti

Quantum key distribution is a cryptographic method that can guarantee secure communication. A satellite-based experiment has shown that this technique can be applied over long distances without the need for trusted relays. **See p.501**

Modern society is driven by the large-scale exchange of information. As a result, secure communication of sensitive data around the world is an increasingly valuable asset. The mathematical toolbox that is widely used for this task can be complemented by applying the principles of quantum physics to enhance the security of the communication link. This approach has highly desirable features, such as protection of the encrypted information from threats that might arise as a consequence of future advances in computational power. However, it also comes with substantial technological challenges in terms of the range of communication possible and the degree of trust in the devices used. Yin *et al.*[1] demonstrate on page 501 that such cryptographic solutions can be deployed over distances exceeding 1,000 kilometres, without compromising the security promised by the underlying quantum technology.

The flagship application of quantum communication is known as quantum key distribution (QKD). This process enables two parties located at a distance from each other to share a secret string of bits (units of information) called a key, which they can use to encrypt and decrypt secret messages, without making assumptions about the computational power of a potential eavesdropper. Although the principle of such absolute security is based solidly on fundamental laws of nature, practical implementations come in different configurations[2].

For example, it is possible for one of the two parties to prepare quantum states of light — the natural physical carrier of information in quantum communication — and to send them to the second party, who measures them. By processing these data using standard classical communication, the two parties can then extract the secret key. QKD in this setting has been demonstrated over 400 km in a low-loss optical fibre[3] and over 1,200 km using a satellite-to-ground communication link[4].

Although impressive, these demonstrations require the two parties' devices to be fully characterized and trusted. Furthermore, losses in the optical-transmission medium eventually become prohibitive. As a result, the networks that need to be established to distribute keys securely between parties contain nodes, which also need to be trusted[5,6]. This constraint might be undesirable for some applications.

If, instead, one could use the distribution of 'entangled' states of light produced by a source, the need for trust would be greatly alleviated. Entangled states embody the peculiar nature of quantum physics and exhibit correlations not found in classical physics. Such correlations can be routed through devices called quantum repeaters, so that remote physical systems can become entangled. The past few years have seen major progress in this direction[7]. But, so far, the longest distances for entanglement distribution have been achieved by transmitting the states directly. These distances are approximately 100 km in an optical fibre[8,9] and 1,200 km using satellite links[10].

Ideally for QKD, the security of the key generated would be confirmed just by detecting these non-classical correlations experimentally, through statistical properties known as Bell inequalities, without having to trust the devices used by the two parties[11]. However, in practice, achieving this level of security places stringent requirements on the experimental devices that cannot be satisfied by currently available technologies. A way forward is to implement entanglement-based QKD that has weaker requirements, whereby, although the parties' devices must be trusted, the source of the entangled states can remain untrusted[12].

Yin *et al.* have performed a complete, long-distance implementation of QKD with these restrictions (Fig. 1). A key way to

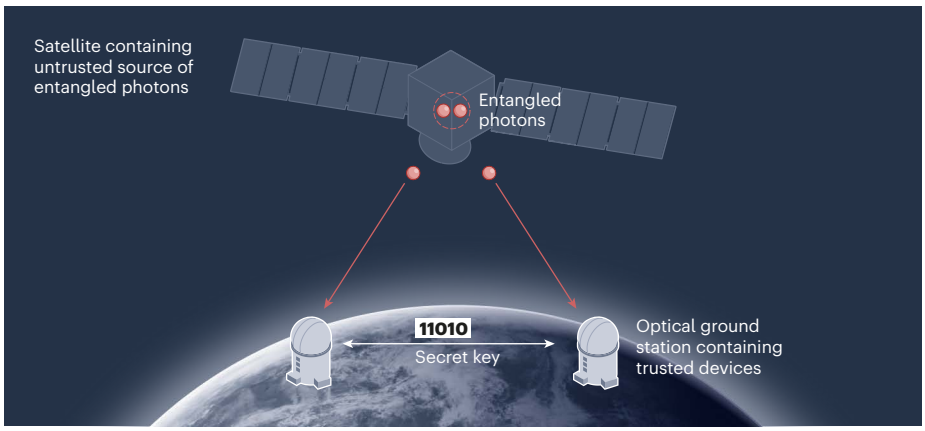


Figure 1 | Entanglement-based quantum cryptography. Yin *et al.*[1] report an experiment in which pairs of entangled photons (photons that are correlated in a non-classical way) are produced on board the satellite Micius. The photons in each pair are then sent to two optical ground stations that are separated by a distance of 1,120 kilometres. This process enables parties at the two stations to share a secret string of bits called a key, which they can use to encrypt and decrypt secret messages with absolute security. In the authors' set-up, the devices used by the two parties must be trusted, but the source of the entangled photons is allowed to be untrusted.

understand their result is to observe how it was achieved, starting from previous work[10] by some of the current authors and their colleagues in 2017. In that paper, the researchers demonstrated the distribution of entangled states generated on board the satellite Micius and sent through two communication links to optical ground stations in China separated by 1,200 km.

Although that work was a milestone for the field, the transmission efficiency achieved was too low for QKD to be carried out in practical conditions. In particular, because only a finite number of states can be transmitted during a short data-collection window, the many errors involved prevented a secret key from being extracted. Taking into account the use of a finite number of transmitted states is crucial for achieving security, especially in the case of a satellite-based experiment, in which data are collected only during the brief time the satellite is visible from the ground stations.

Yin *et al.* remedied this problem by implementing major technological enhancements. These included installing highly efficient telescopes at the ground stations and optimizing equipment components at all stages of the optical path. The authors' meticulous optimization also involved cutting-edge signal acquisition, pointing and tracking systems and synchronization techniques for both the satellite and the ground stations. Their efforts led to a fourfold increase in transmission efficiency compared with the previous experiment and, consequently, produced low enough error rates for a secret key to be extracted. The authors also verified the stability and reliability of their findings over multiple satellite orbits.

From a security perspective, this demonstration does not remove the need for trust in the receiving stations. Therefore, assumptions must be made about the internal workings of the devices in these stations. Yin *et al.* did two things to minimize the risk that these assumptions would not hold in practice. First, they used a systematic approach to tackling imperfections that might inadvertently leak information to a potential eavesdropper. Second, they used a range of solutions to actively control the properties of the photonic information carriers. Combined with security from this quantum approach that should be guaranteed against all possible attacks, this makes the authors' result the most advanced QKD demonstration so far.

However, several shortcomings will need to be overcome for these findings to become relevant for truly practical high-security applications. For instance, the experiment produced keys at extremely low rates. Also, the experiment was carried out only at night, and using a wavelength that is incompatible with the optical-fibre networks used for telecommunication that would interface with space-based networks in infrastructures for global quantum communication. Moreover, QKD can be achieved only between ground stations that are visible simultaneously from the satellite.

Progress in all these areas requires the development of high-performance devices operating at a longer wavelength than that used in this work, the use of satellites in higher orbits than that of Micius and – in the long term – integration of the demonstrated technology with quantum repeaters and other promising architectures allowing for untrusted nodes[13]. Such advances would then unlock the full potential of quantum technologies for executing cryptographic tasks at a global scale.

Eleni Diamanti is at the National Centre for Scientific Research (CNRS), Sorbonne University, 75005 Paris, France.
e-mail: eleni.diamanti@lip6.fr

1. Yin, J. *et al. Nature* **582**, 501–505 (2020).
2. Diamanti, E., Lo, H.-K., Qi, B. & Yuan, Z. *npj Quantum Inf.* **2**, 16025 (2016).
3. Boaron, A. *et al. Phys. Rev. Lett.* **121**, 190502 (2018).
4. Liao, S.-K. *et al. Nature* **549**, 43–47 (2017).
5. Peev, M. *et al. N. J. Phys.* **11**, 075001 (2009).
6. Liao, S.-K. *et al. Phys. Rev. Lett.* **120**, 030501 (2018).
7. Bhaskar, M. K. *et al. Nature* **580**, 60–64 (2020).
8. Aktas, D. *et al. Laser Photon. Rev.* **10**, 451–457 (2016).
9. Wengerowsky, S. *et al. Proc. Natl Acad. Sci. USA* **116**, 6684–6688 (2019).
10. Yin, J. *et al. Science* **356**, 1140–1144 (2017).
11. Ekert, A. K. *Phys. Rev. Lett.* **67**, 661–663 (1991).
12. Bennett, C. H., Brassard, G. & Mermin, N. D. *Phys. Rev. Lett.* **68**, 557–559 (1992).
13. Lo, H.-K., Curty, M. & Qi, B. *Phys. Rev. Lett.* **108**, 130503 (2012).

Ecology

Model system might reveal how coral cells evict algae

Alejandro Sánchez Alvarado

Global warming is threatening the survival of coral reefs. A laboratory model system has now been developed that should aid efforts to understand reef biology and the processes that underlie harmful bleaching events. **See p.534**

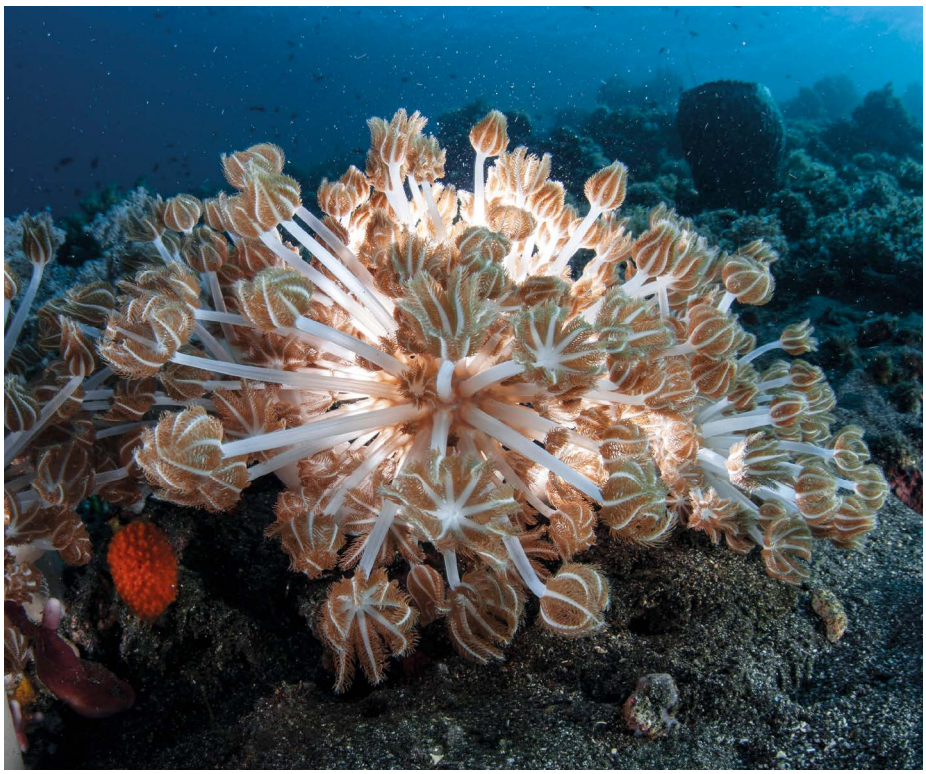
From the late Precambrian era (around 650 million years ago) to the present day, a singular, carbon-based 'economy' has been operating between corals and algae that has fuelled the building of untold expanses of barrier reefs in oceans around the globe. On page 534, Hu *et al.*[1] now set the stage for efforts to gain a deeper understanding of how corals and algae interact in coral reefs.

Corals – multicellular marine invertebrates belonging to the class Anthozoa of the phylum Cnidaria – usually live in compact colonies composed of individual structures called polyps. Most reef-building corals harbour algae in their cells in a specialized, membrane-bound compartment called a symbiosome. As its name implies, this specialized structure is home to one of nature's most remarkable, mutually beneficial, endosymbiotic relationships. Corals provide specific species of alga with a protected environment and with compounds needed to carry out photosynthesis. In return, the algae supply the coral with the products of photosynthesis: oxygen, glucose, glycerol and amino acids. This biomolecular bounty is then transformed by the corals into proteins, fats, carbohydrates and a calcium carbonate skeleton.

> **"The authors confirmed that they had correctly identified the coral cells that host algae."**

Around 90% of the organic material produced by these algal endosymbionts is ultimately transferred to the coral host[2], underpinning the quiet yet unceasing growth and productivity of coral reefs[3]. It has been estimated that this endosymbiotic relationship is responsible for an area of nearly 250,000 square kilometres of the most spectacular and crucial ecosystems on our planet, supporting some 2 million or more species[4]. However, the current rise in ocean temperatures is causing disruption, because exposure to prolonged heat causes corals to evict their symbiotic algae, resulting in a phenomenon known as coral bleaching – loss of the colourful algae leaves the coral white in appearance. Severe coral bleaching threatens to cause a marine calamity of global proportions. Unfortunately, we know little about the molecular basis that underlies how coral cells orchestrate algal expulsion, nor about how corals recognize, take up and maintain their algal endosymbionts.

Hu and colleagues' work heralds a new



ETHAN DANIELS/SPL

Figure 1 | *Xenia*, a fast-growing species of coral found in the Red Sea and Indian Ocean.

and valuable attempt at a rigorous and systematic understanding of the cellular and molecular basis of coral endosymbiosis, and, potentially, of coral bleaching. The authors selected a fast-growing species of *Xenia* coral (Fig. 1) that usually resides in waters stretching from the Red Sea to the Indian Ocean. *Xenia* establishes an endosymbiotic relationship with algae of the Symbiodiniaceae family. Hu *et al.* assembled a genome sequence for *Xenia*, complete with chromosome-level information. The authors also generated a map down to a resolution of single cells, revealing the RNA profiles of the cells present in a *Xenia* polyp. This approach has effectively brought the power of both genomics and bioinformatics to the study of coral biology. The authors report that the *Xenia* genome has approximately 24,000 genes, and the single-cell atlas reveals 16 cell clusters, each of which has a distinct gene-expression pattern.

Using this information, the authors sought to resolve a key mystery in coral biology: which cells in the organism are responsible for recognizing the appropriate algal species and establishing the endosymbiosis? By taking advantage of the visible autofluorescence of the algal partner, Hu *et al.* used a flow-cytometry approach to separate the alga-containing from the alga-free coral cells. RNA sequencing then enabled the authors to determine which genes were expressed in the two cell populations. By comparing this information with the RNA-sequencing data from their single-cell atlas, Hu *et al.* found that cluster 16 showed the highest overall similarity to the profile of the alga-containing cells. The authors confirmed that they had correctly identified the coral cells that host algae *in vivo* by using a technique called *in situ* hybridization to detect the expression of the cluster-16 genes associated with the proposed host cell. Remarkably, the authors found that these proposed endosymbiotic cells corresponded to a mere 1.4% of all the coral cells catalogued in their single-cell atlas.

Because single-cell RNA studies normally capture a single moment in time, the developmental provenance of the endosymbiotic cells in cluster 16 remained unclear. The authors addressed this concern by resorting to 'development on demand' — the regeneration of missing body parts after cellular 'amputation'. By surgically removing all the tentacles from *Xenia* polyps, restoration of the endosymbiotic cells could be followed from scratch. Using a combination of single-cell RNA sequencing, analysis of gene-expression patterns and a 'pulse–chase' method to label and track cells, the authors describe a lineage for endosymbiotic cells that progresses from progenitor cells to alga-uptake cells, and from mature alga-containing cells to cells devoid of algae. Hu and colleagues provide examples of differentially expressed genes that serve to mark each of the lineage stages and that might give insights into the molecular machinery driving the activities of coral endosymbiotic cells.

Nevertheless, Hu and colleagues' findings await detailed functional validation. Cause-and-effect relationships will need to be determined for the many molecules identified and for the cellular activities associated with algal recognition, uptake, maintenance and eviction. This should be possible in the near future by using methods to reduce the expression of targeted genes through the introduction of artificial molecules called short hairpin RNAs — either by microinjection or through an electroporation method used in another cnidarian, the starlet sea anemone *Nematostella vectensis*[5]. Alternatively, permanent modifications of the genome might be desirable, and thus might require harnessing the CRISPR gene-editing technique either to introduce mutations or to add desired sequences. This prospect should be helped by the high-quality *Xenia* genome that the authors have made available.

Future studies should focus on the cell biology of endosymbiosis. Visualizing the lineage progression from progenitor to endosymbiotic cell should reveal fascinating aspects of cell biology. For example, it could help us to understand the mechanisms by which coral cells expand to allow them to take up algae that are similar in size to themselves.

Coral reefs provide the foundation for an enduring and evolving legacy of progress in knowledge about aspects of their biology, even though many features of their existence are still poorly understood. The development of a laboratory-friendly, coral-research organism suitable for molecular and cellular experimentation is of great significance, and the importance of this achievement cannot be overemphasized. Hu and colleagues' work opens the door to the possibility of identifying the principles by which corals recognize, take up and expel their endosymbionts. Understanding the mechanisms underlying any of these processes will not only enhance our understanding of symbiosis, but also contribute to the study and, more crucially, to the development of possible solutions to one of the major problems affecting the health of our planet.

Alejandro Sánchez Alvarado is at the Stowers Institute for Medical Research and at the Howard Hughes Medical Institute, Kansas City, Missouri 64110, USA.
e-mail: asa@stowers.org

1. Hu, M., Zheng, X., Fan, C.-M. & Zheng, Y. *Nature* **582**, 534–538 (2020).
2. Sumich, J. L. *An Introduction to the Biology of Marine Life* 6th edn, Ch. 9, 255–269 (Brown, 1996).
3. Levinton, J. S. *Marine Biology: Function, Biodiversity, Ecology* Ch. 14, 306–319 (Oxford Univ. Press, 1995).
4. Spalding, M. D., Ravilious, C. & Green, E. P. *World Atlas of Coral Reefs* Ch. 1, 9–45 (Univ. California Press, 2001).
5. Karabulut, A., He, S., Chen, C. Y., McKinney, S. A. & Gibson, M. C. *Dev. Biol.* **448**, 7–15 (2019).

This article was published online on 17 June 2020.

A planet within the debris disk around the pre-main-sequence star AU Microscopii

https://doi.org/10.1038/s41586-020-2400-z

Received: 16 February 2019

Accepted: 17 March 2020

Published online: 24 June 2020

Check for updates

Peter Plavchan[1]✉, Thomas Barclay[2,13], Jonathan Gagné[3], Peter Gao[4], Bryson Cale[1], William Matzko[1], Diana Dragomir[5,6], Sam Quinn[7], Dax Feliz[8], Keivan Stassun[8], Ian J. M. Crossfield[5,9], David A. Berardo[5], David W. Latham[7], Ben Tieu[1], Guillem Anglada-Escudé[10], George Ricker[5], Roland Vanderspek[5], Sara Seager[5], Joshua N. Winn[11], Jon M. Jenkins[12], Stephen Rinehart[13], Akshata Krishnamurthy[5], Scott Dynes[5], John Doty[13], Fred Adams[14], Dennis A. Afanasev[13], Chas Beichman[15,16], Mike Bottom[17], Brendan P. Bowler[18], Carolyn Brinkworth[19], Carolyn J. Brown[20], Andrew Cancino[21], David R. Ciardi[16], Mark Clampin[13], Jake T. Clark[20], Karen Collins[7], Cassy Davison[22], Daniel Foreman-Mackey[23], Elise Furlan[15], Eric J. Gaidos[24], Claire Geneser[25], Frank Giddens[21], Emily Gilbert[26], Ryan Hall[22], Coel Hellier[27], Todd Henry[28], Jonathan Horner[20], Andrew W. Howard[29], Chelsea Huang[5], Joseph Huber[21], Stephen R. Kane[30], Matthew Kenworthy[31], John Kielkopf[32], David Kipping[33], Chris Klenke[21], Ethan Kruse[13], Natasha Latouf[1], Patrick Lowrance[34], Bertrand Mennesson[15], Matthew Mengel[20], Sean M. Mills[29], Tim Morton[35], Norio Narita[36,37,38,39,40], Elisabeth Newton[41], America Nishimoto[21], Jack Okumura[20], Enric Palle[40], Joshua Pepper[42], Elisa V. Quintana[13], Aki Roberge[13], Veronica Roccatagliata[43,44,45], Joshua E. Schlieder[13], Angelle Tanner[25], Johanna Teske[46], C. G. Tinney[47], Andrew Vanderburg[18], Kaspar von Braun[48], Bernie Walp[49], Jason Wang[4,29], Sharon Xuesong Wang[46], Denise Weigand[21], Russel White[22], Robert A. Wittenmyer[20], Duncan J. Wright[20], Allison Youngblood[13], Hui Zhang[50] & Perri Zilberman[51]

AU Microscopii (AU Mic) is the second closest pre-main-sequence star, at a distance of 9.79 parsecs and with an age of 22 million years[1]. AU Mic possesses a relatively rare[2] and spatially resolved[3] edge-on debris disk extending from about 35 to 210 astronomical units from the star[4], and with clumps exhibiting non-Keplerian motion[5–7]. Detection of newly formed planets around such a star is challenged by the presence of spots, plage, flares and other manifestations of magnetic 'activity' on the star[8,9]. Here we report observations of a planet transiting AU Mic. The transiting planet, AU Mic b, has an orbital period of 8.46 days, an orbital distance of 0.07 astronomical units, a radius of 0.4 Jupiter radii, and a mass of less than 0.18 Jupiter masses at 3σ confidence. Our observations of a planet co-existing with a debris disk offer the opportunity to test the predictions of current models of planet formation and evolution.

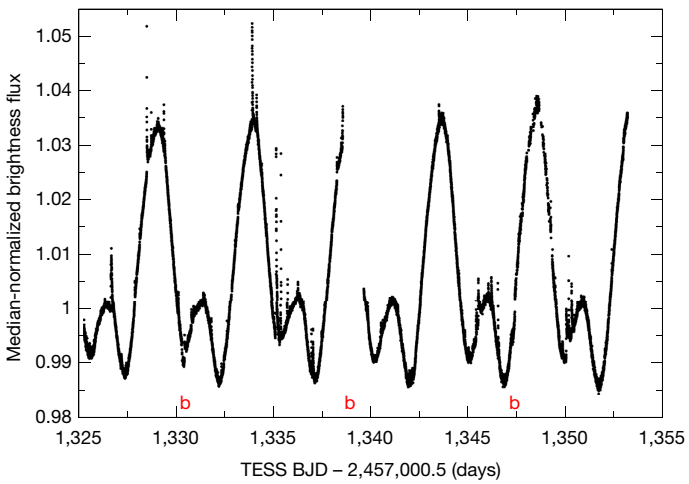


Fig. 1 | TESS light curve for AU Mic. Black dots, normalized flux as a function of time, obtained from the MAST archive. Transit ephemerides of AU Mic b are indicated as 'b' in red. The double-humped sinusoidal-like pattern is due to the rotational modulation of starspots, with the 4.863-day rotation period readily apparent. The large, brief vertical streaks of data points deviating upwards from this slower modulation are due to flares. Data with non-zero quality flags indicating the presence of spacecraft-related artefacts, such as momentum dumps (see Fig. 2 legend), are removed. The gap at about 1,339 days corresponds to a gap in the data downlink with Earth during the spacecraft's perigee. A third transit of AU Mic b was missed during this data downlink data gap, and thus the orbital period of AU Mic b is one-half of the period inferred from the two TESS transit events seen. AU Mic exhibited flaring activity with energies ranging from $10^{31.6}$ to $10^{33.7}$ erg in the TESS bandpass over the 27-day light curve (±~60%), with a mean flare amplitude of 0.01 relative flux units. 1σ measurement uncertainties are smaller than the symbols shown (<1 parts per thousand, p.p.t.).

A list of affiliations appears at the end of the paper.

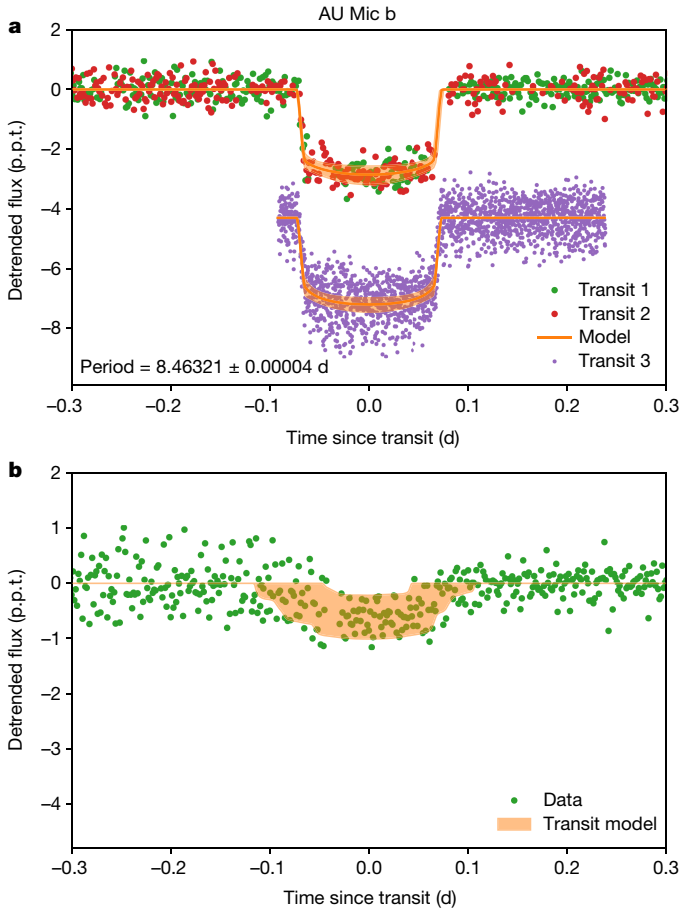


Fig. 2 | Light curves of the transits of AU Mic b, and a separate, candidate transit event. a, Data points show light curves from TESS in visible light (green and red filled circles for transits 1 and 2, respectively) and from Spitzer IRAC[11] at 4.5 μm wavelength (purple filled circles for transit 3). The data for transits of AU Mic b are shown with an arbitrary vertical shift applied for clarity; flux units are p.p.t. The transit model (orange curve) includes a photometric model that accounts for the stellar activity modelled with a Gaussian Process (GP), which is subtracted from the data before plotting. The frequent flares from the stellar surface are removed with an iterative sigma-clipping (see Methods). In particular, flares are observed during the egress of both the TESS transits of AU Mic b, and also just after the ingress of the second transit of AU Mic b. The presence of these flares in the light curve particularly affect our precision in measuring the transit duration and thus the mass/density of the host star AU Mic, and consequently the impact parameter and eccentricity of the orbit of AU Mic b. Model uncertainties shown as shaded regions are 1σ confidence intervals. The uncertainty in the out-of-transit baseline is about 0.5 p.p.t. but is not shown for clarity. **b**, The AU Mic candidate single transit signal, identified by visual inspection of the TESS light curve. The change in noise before and after the candidate transit signal is due to a 'dump' of angular momentum from the spacecraft reaction wheels which decreased the pointing jitter and improved the photometric precision; data points during the dump are not shown.

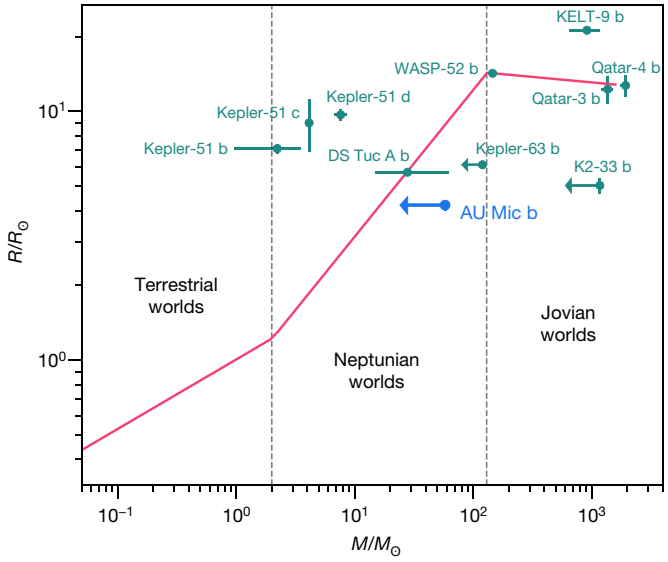


Fig. 3 | Mass–radius diagram showing AU Mic b in the context of 'mature' exoplanets and known young exoplanets. Mass M and radius R are normalized to the values for Earth, respectively $M_\oplus$ and $R_\oplus$. AU Mic b is shown in blue; we compare it to the nominal best-fit mass–radius relationship from known exoplanets orbiting older main-sequence stars[19], shown as a red segmented line (dispersion not shown), and known exoplanets from the NASA Exoplanet Archive with measured masses or mass upper limits, radii, and estimated stellar host ages ≤400 Myr, as follows: DS Tuc A b (mass is estimated from ref. [19] and not measured), Kepler-51 bcd, Kepler 63 b, K2-33 b, Qatar-3 b, Qatar-4 b, KELT-9 and WASP-52 b. By combining the radius measurement from TESS, and the mass upper limit from radial velocities (RVs), we can ascertain an upper limit to the density of AU Mic b to critically inform models for planet formation. Our current upper limit for the mass of AU Mic b cannot rule out a density consistent with Neptune-like planets orbiting older main-sequence stars, but a more precise constraint or measurement in the future may show it to be inflated. Uncertainties shown are 1σ for detections, and 3σ for mass upper limits.

NASA's Transiting Exoplanet Survey Satellite (TESS) mission[10] was launched on 18 April 2018, and monitored the brightness of AU Mic during the first 27 days of its survey of most of the sky (Fig. 1). Two transits of AU Mic b appear in the TESS photometric light curve. Follow-up observations with the Spitzer Space Telescope[11] confirm the transits of AU Mic b. Our analyses show that this transiting planet has an orbital period of 8.46 days, an orbital distance of 0.07 astronomical units (AU) and a radius of 0.4 Jupiter radii. An additional, shallower candidate transit is observed in the TESS light curve, which suggests the possible existence of additional planets (Fig. 2). Joint radial-velocity (RV) and high-resolution adaptive optics imaging rules out[12] other planets in this system more massive than Jupiter interior to about 20 AU. The 3σ upper limit to the velocity reflex motion semi-amplitude for AU Mic b is $K < 28$ m s^{-1} (see Methods), corresponding to an upper limit for the mass of AU Mic b of <0.18 Jupiter masses ($M_{\rm Jupiter}$) or <3.4 Neptune masses ($M_{\rm Neptune}$; see Fig. 3, Tables 1 and 2).

The proximity, brightness, age and edge-on geometry of the AU Mic system will permit us to study AU Mic b at an early stage of its dynamical, thermal and atmospheric evolution, as well as any connection between the planet and the residual debris disk. The host star is a red dwarf, one of the most abundant stellar types in our Galaxy. Their diminutive size, mass and luminosity make middle-aged, comparatively inactive M dwarfs favoured targets to search for Earth-size planets in circumstellar habitable zones. Thus AU Mic is an opportunity to study a possible antecedent to these important systems. Moreover, AU Mic, unlike most M dwarfs of a similar age, possesses a debris disk[2], and hence may offer insight into connections between planets and dust disks. This system confirms[13] that gaseous planet formation and any primordial disk migration takes place in less than 20 Myr. The accretion and migration of this (or additional) planets could have left behind the Kuiper-belt-like 'birth ring' of parent body debris that is hypothesized[6] at about 35 AU, while clearing the interior disk of gas and dust. Furthermore, it is possible that any remnant primordial debris in the inner disk near the current locations of the planet could be in the process of being ejected by this planet. Measurement of the spin-orbit obliquity of AU Mic b via the Rossiter–McLaughlin effect (a peak-to-peak amplitude of 40 m s^{-1} is expected) or Doppler tomography would be

Table 1 | Star parameters

Parameter	68% credible interval	Remarks
AU Mic (star)		
Distance from the Sun	9.79 ± 0.04 pc	Gaia mission parallax
Radius	$(0.75 \pm 0.03)R_{\odot}$	Directly measured with interferometry[17]
Mass	$(0.50 \pm 0.03)M_{\odot}$	Estimated from spectral type and age[a]
$T_{\rm eff}$	3,700 ± 100 K	Spectral energy distribution modelling[15]
Luminosity	$0.09L_{\odot}$	Spectral energy distribution modelling[15]
Age	22 ± 3 Myr	Ref. [1]
Rotation period	4.863 ± 0.010 days	RV analysis, TESS light curve, SuperWASP light curve[18]
Projected rotational velocity	8.7 ± 0.2 km s^{-1}	Ref. [12]
Linear limb-darkening coefficient (TESS)	$0.21^{+0.20}_{-0.15}$	TESS light curve
Quadratic limb-darkening coefficient (TESS)	$0.0^{+0.18}_{-0.14}$	TESS light curve
Linear limb-darkening coefficient (Spitzer)	$0.17^{0.22}_{-0.12}$	Spitzer light curve
Quadratic limb-darkening coefficient (Spitzer)	$0.15^{+0.27}_{-0.21}$	Spitzer light curve
Visible stellar activity amplitude	145^{+17}_{-14} ms^{-1}	RV analysis
Near-infrared stellar activity amplitude	80^{+16}_{-12} ms^{-1}	RV analysis; K band at 2.3 μm
Spot decay half-life	110 ± 30 days	RV analysis
GP hyper-parameter 4	0.37 ± 0.02	RV analysis
Apparent magnitude	TESS = 6.76 mag	TESS light curve

[a]Also consistent with independently fitting the two transit events in TESS light curve for AU Mic b.

Table 2 | Planetary parameters

Parameter	68% credible interval	Remarks
AU Mic b		
Period	8.46321 ± 0.00004 days	TESS and Spitzer transit light curve analysis
Semi-major axis	$0.066^{+0.007}_{-0.006}$ AU	TESS and Spitzer transit light curve analysis
Velocity semi-amplitude, K	<28 m s^{-1}	RV analysis
Mass	$<3.4M_{\rm Neptune}$ $<0.18M_{\rm Jupiter}$	RV analysis
Radius	$(1.08 \pm 0.05)R_{\rm Neptune}$ $(0.375 \pm 0.018)R_{\rm Jupiter}$	TESS and Spitzer transit light curve
Density	<4.4 g cm^{-3}	RV / TESS analysis
Time(s) of conjunction	$2,458,330.39153^{+0.00070}_{-0.00068}$ BJD[a]	TESS and Spitzer transit light curves
Transit duration, τ_{14}	$3.50^{+0.63}_{-0.59}$ h	TESS and Spitzer transit light curves
$R_{\rm p}/R_{*}$	0.0514 ± 0.0013	TESS and Spitzer transit light curve
Impact parameter, b	$0.16^{+0.14}_{-0.11}$	TESS and Spitzer transit light curve
a/R_{*}	$19.1^{+1.8}_{-1.6}$	TESS and Spitzer transit light curve
Eccentricity	$0.10^{+0.17}_{-0.09}$	TESS and Spitzer transit light curve[b].
Candidate transit event		
Period	30 ± 6 days	TESS light curve transit duration
Radius	$(0.60 \pm 0.17)R_{\rm Neptune}$ = $(0.21 \pm 0.06)R_{\rm Jupiter}$	TESS transit light curve
Time(s) of conjunction	2,458,342.22 ± 0.03 days	TESS transit light curve
$R_{\rm p}/R_{*}$	0.028 ± 0.006	TESS transit light curve
Impact parameter, b	0.5 ± 0.3	TESS transit light curve
a/R_{*}	40 ± 8	TESS transit light curve
Eccentricity	0.2 ± 0.2	TESS transit light curve

[a]Barycentric Julian Day.
[b]Circular orbit assumed for RV analysis.

an important test of migration models since we expect any obliquity in this young system to be unaffected by stellar tides and thus primordial.

AU Mic is a member of the β Pictoris Moving Group; the group's archetype β Pic is a much more massive (about 3.5×), luminous (about 100×) and hotter (approximately 2×) A-type star, also possessing a debris disk. β Pic has a more massive Jovian planet β Pic b observed by direct imaging at a semi-major axis of about 9 AU, with a mass of approximately $(11\pm2)M_{\rm Jupiter}$ determined with astrometry[14]. AU Mic and β Pic are of the same stellar age, but are very different exoplanet host stars. While AU Mic b possibly formed at a distance similar to β Pic b and then migrated inwards to its present location, β Pic b has not substantially migrated inward. These two coeval systems provide an excellent differential comparison for planet formation.

Finally, the combined effect of stellar winds and interior planets have been invoked to explain the high-speed ejection of dust clumps from the system[6,7]. The observed clumps are dynamically decoupled from AU Mic b; the ratio of the semi-major axes (0.06 AU versus >35 AU) is >100, but the clumps could have originated much closer to the star. Dust produced in the debris ring at about 35 AU will spiral inwards primarily as a result of stellar wind drag, which, for AU Mic and a mass loss rate about 1,000 times that of the solar wind[6], is estimated to be 3,700 times stronger than Poynting–Robertson drag[2]. To compare the timescales between collisions of dusty debris and the stellar wind drag force[15], we assume a birth ring fractional width of 10% (3.5 AU), and given AU Mic's infrared flux excess, find that the stellar wind drag and dust collision timescales are roughly equal. Thus, some fraction of the dust grains generated in the birth ring at about 35 AU may spiral inward to the host star under the action of stellar wind drag, instead of being ground down further by dust collisions until blown out of the system by radiation pressure. For 1-μm-sized solid grains of dusty debris, the in-spiral time would be approximately 7,500 years, much shorter than the age of the star. Such dust may have been observed by ALMA[16] at <3 AU, interior to the birth ring at 35 AU. Dust reaching the orbit of an interior planet could be dynamically ejected, depending on the Safronov number: we estimate that of AU Mic b to be 0.07 and thus inefficient at ejecting dust.

There is no other known system that possesses all of these crucial pieces—an M-dwarf star that is young, nearby, still surrounded by a debris disk within which are moving clumps, and orbited by a planet with a direct radius measurement. As such, AU Mic provides a unique laboratory to study and model planet and planetary atmosphere formation and evolution processes in detail.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2400-z.

1. Mamajek, E. E. & Bell, C. P. M. On the age of the β Pictoris moving group. *Mon. Not. R. Astron. Soc.* **445**, 2169–2180 (2014).
2. Plavchan, P., Jura, M. & Lipscy, S. J. Where are the M dwarf disks older than 10 million years? *Astrophys. J.* **631**, 1161 (2005).
3. Kalas, P., Liu, M. C. & Matthews, B. C. Discovery of a large dust disk around the nearby star AU Microscopii. *Science* **303**, 1990–1992 (2004).
4. Strubbe, L. E. & Chiang, E. I. Dust dynamics, surface brightness profiles, and thermal spectra of debris disks: the case of AU Microscopii. *Astrophys. J.* **648**, 652 (2006).
5. Boccaletti, A. et al. Fast-moving features in the debris disk around AU Microscopii. *Nature* **526**, 230–232 (2015).
6. Chiang, E. & Fung, J. Stellar winds and dust avalanches in the AU Mic debris disk. *Astrophys. J.* **848**, 4 (2017)
7. Sezestre, É. et al. Expelled grains from an unseen parent body around AU Microscopii. *Astron. Astrophys.* **607**, A65 (2017).
8. van Eyken, J. et al. The PTF Orion project: a possible planet transiting a T-Tauri star. *Astrophys. J.* **755**, 42 (2012).
9. Donati, J. F. et al. A hot Jupiter orbiting a 2-million-year-old solar-mass T Tauri star. *Nature* **534**, 662–666 (2016).
10. Ricker, G. R. et al. Transiting Exoplanet Survey Satellite (TESS). *J. Astron. Telesc. Instrum. Syst.* **1**, 014003 (2014).
11. Deming, D. et al. Spitzer secondary eclipses of the dense, modestly-irradiated, giant exoplanet HAT-P-20b using pixel-level decorrelation. *Astrophys. J.* **805**, 132 (2015).
12. Lannier, J. et al. Combining direct imaging and radial velocity data towards a full exploration of the giant planet population. I. Method and first results. *Astron. Astrophys.* **603**, A54 (2017).
13. Kley, W. & Nelson, R. P. Planet-disk interaction and orbital evolution. *Annu. Rev. Astron. Astrophys.* **50**, 211–249 (2012).
14. Snellen, I. A. G. & Brown, A. G. A. The mass of the young planet Beta Pictoris b through the astrometric motion of its host star. *Nature Astron.* **2**, 883–886 (2018).
15. Plavchan, P. et al. New debris disks around young, low-mass stars discovered with the Spitzer Space Telescope. *Astrophys. J.* **698**, 1068–1094 (2009).
16. MacGregor, M. A. et al. Millimeter emission structure in the first ALMA image of the AU Mic debris disk. *Astrophys. J.* **762**, L21 (2013).
17. White, R. et al. Stellar radius measurements of the young debris disk host AU Mic. *Proc. AAS Meet.* **233**, 348.12 (2015).
18. Torres, C. A. O., Ferraz Mello, S. & Quast, G. R. HD 197481: a periodic dMe variable star. *Astrophys. J.* **11**, L13–L14 (1972).
19. Chen, J. & Kipping, D. Probabilistic forecasting of the masses and radii of other worlds. *Astrophys. J.* **834**, 17 (2017).

[1]Department of Physics and Astronomy, George Mason University, Fairfax, VA, USA. [2]Center for Space Sciences and Technology, University of Maryland Baltimore County (UMBC), Baltimore, MD, USA. [3]Institute for Research on Exoplanets, Département de Physique, Université de Montréal, Montréal, Quebec, Canada. [4]Department of Astronomy, University of California, Berkeley, CA, USA. [5]Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Department of Physics and Astronomy, University of New Mexico, Albuquerque, NM, USA. [7]Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA. [8]Department of Physics and Astronomy, Vanderbilt University, Nashville, TN, USA. [9]Department of Physics and Astronomy, University of Kansas, Lawrence, KS, USA. [10]School of Physics and Astronomy, Queen Mary, University of London, London, UK. [11]Department of Astrophysical Sciences, Princeton University, Princeton, NJ, USA. [12]SETI Institute, Mountain View, CA, USA. [13]Exoplanets and Stellar Astrophysics Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, USA. [14]Department of Astronomy, University of Michigan, Ann Arbor, MI, USA. [15]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA. [16]NASA Exoplanet Science Institute, California Institute of Technology, Pasadena, CA, USA. [17]Institute for Astronomy, University of Hawaii at Manoa, Honolulu, HI, USA. [18]Department of Astronomy, University of Texas at Austin, Austin, TX, USA. [19]University Corporation for Atmospheric Research, Boulder, CO, USA. [20]University of Southern Queensland, Centre for Astrophysics, Toowoomba, Queensland, Australia. [21]Department of Physics, Astronomy and Materials Science, Missouri State University, Springfield, MO, USA. [22]Department of Physics and Astronomy, Georgia State University, Atlanta, GA, USA. [23]Center for Computational Astrophysics, Flatiron Institute, New York, NY, USA. [24]Department of Earth Sciences, University of Hawaii at Manoa, Honolulu, HI, USA. [25]Department of Physics and Astronomy, Mississippi State University, Starkville, MS, USA. [26]Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL, USA. [27]Keele University, Keele, Staffordshire, UK. [28]RECONS Institute, Chambersburg, PA, USA. [29]Department of Astronomy, California Institute of Technology, Pasadena, CA, USA. [30]Department of Earth and Planetary Sciences, University of California, Riverside, CA, USA. [31]Leiden Observatory, Leiden University, Leiden, The Netherlands. [32]Department of Physics and Astronomy, University of Louisville, Louisville, KY, USA. [33]Department of Astronomy, Columbia University, New York, NY, USA. [34]IPAC, California Institute of Technology, Pasadena, CA, USA. [35]Astronomy Department, University of Florida, Gainesville, FL, USA. [36]Department of Astronomy, The University of Tokyo, Tokyo, Japan. [37]JST, PRESTO, Tokyo, Japan. [38]Astrobiology Center, NINS, Tokyo, Japan. [39]National Astronomical Observatory of Japan, NINS, Tokyo, Japan. [40]Instituto de Astrofisica de Canarias (IAC), La Laguna, Tenerife, Spain. [41]Department of Physics and Astronomy, Dartmouth College, Hanover, NH, USA. [42]Department of Physics, Lehigh University, Bethlehem, PA, USA. [43]Dipartimento di Fisica "Enrico Fermi", Universita' di Pisa, Pisa, Italy. [44]INAF-Osservatorio Astrofisico di Arcetri, Firenze, Italy. [45]INFN, Sezione di Pisa, Pisa, Italy. [46]Observatories of the Carnegie Institution for Science, Pasadena, CA, USA. [47]Exoplanetary Science at UNSW, School of Physics, UNSW Sydney, New South Wales, Australia. [48]Lowell Observatory, Flagstaff, AZ, USA. [49]NASA Infrared Telescope Facility, Hilo, HI, USA. [50]School of Astronomy and Space Science, Key Laboratory of Ministry of Education, Nanjing University, Nanjing, China. [51]SUNY Stony Brook, Stony Brook, NY, USA. ✉e-mail: pplavcha@gmu.edu

Methods

TESS light-curve analysis

AU Mic has long been known as a young star exhibiting flares and brightness variations driven by large starspots on the stellar surface rotating in and out of view[20]. Previous attempts to find transiting planets were not successful owing to this variability and the redness of the star combined with secondary atmospheric extinction effects[21,22], in spite of reasoning that the orbits of any planets could be aligned with AU Mic's edge-on debris disk, and therefore could be more likely to transit than for a random inclination.

TESS observed AU Mic (TIC 441420236) in its first sector (2018 July 25–August 22). The TESS light curve from the 2-min cadence stamp was processed by the Science Processing Operations Center pipeline, a descendant of the Kepler mission pipeline based at the NASA Ames Research Center[23,24]. After visually identifying the transits in the light curve, we independently validate the existence of the transits from the 30-min full-frame image (FFI) data. We also extract light curves with different photometric apertures, and confirm that the transit signal is robust and consistent. No centroid motion is observed during transits, suggesting that it is associated with AU Mic rather than being an instrumental systematic or contamination from scattered background light or a distant star. To validate the transit with ancillary data, we inspect archival sky survey images such as POSS and find no background stars within the TESS pixels that are present at the location of AU Mic with a sufficient brightness ratio so as to mimic the observed transit signals with a background eclipsing binary. Nor do we or others identify any background eclipsing binaries in high-contrast adaptive optics imaging[3] or our high-resolution spectroscopy (see below). The nearest Gaia DR2 source that is capable of producing a false positive if an eclipsing binary (with G-band contrast = 5.7 mag, ignoring TESS-G-band colour terms) is 76 arcsec or 3 TESS pixels from AU Mic. Finally, the interferometric stellar radius determination[17] rules out bound stellar companions.

We perform multiple independent analyses of the TESS light curve to identify and model the transits present, including the TESS mission pipeline planet detection algorithms, ExoFAST v1.0 and v2.0[25,26], and asterodensity profiling[27], which yield consistent results. While ExoFAST does support the simultaneous modelling of light curves and RVs, it does not include components for modelling the stellar activity prevalent for AU Mic in the RVs. Thus, we carry out independent analyses of the light curves and RVs. For the TESS light curve, ExoFAST and astrodensity profiling do not simultaneously model the exoplanet transits and detrending of the photometric variability produced by the rotational modulation of the starspots. Thus to prepare the TESS light curve for these analysis tools, we first fit four sinusoids to the light curve with periods equal to the rotation period, and one-half, one-third and one-quarter thereof. We then apply a 401 data-point running median filter to remove the remaining photometric modulation due to starspots. The flares present in the transit events were not removed for these analyses, primarily affecting the determination of the transit duration of AU Mic b.

Spitzer light-curve analysis

Owing to the data collection gap in the TESS light curve, Spitzer Director's Discretionary Time (DDT; Program ID no. 14214, 17.3 h time allocation) observations were proposed, awarded and collected in 2019 to rule in or rule out one-half of the orbit period for AU Mic b as seen in the TESS light curve. Three transits were observed with IRAC at 4.5 μm, one of which is presented herein, the others will be presented in a future paper. We first clean up the raw images by sigma-clipping outliers and subtracting off a background estimate from an annulus around the centre of light. We then sum the flux in a circular aperture centred around the centre of light of each frame, and do this for several different aperture radii. We then follow the procedure from ref. [11] and do a pixel level decorrelation (PLD; using 3 × 3 pixels) on each radius, and pick the one that gives the smallest scatter. We adopt a 2.4 pixel radius aperture, binned by a factor of 106.

Joint TESS and Spitzer photometric analysis

We carry out a custom analysis that simultaneously accounts for the rotational modulation of starspots, the flares and the transit events for both the TESS and Spitzer light curves to evaluate the impact our detrending of the spot rotational modulation and flares has on our analysis of the transit events: this is the analysis we adopt in the main text (Extended Data Fig. 1). We use the TESS pre-search data conditioned light curve created by the TESS pipeline[24,28,29] for this analysis. To remove flares, we create a smoothed version of the light curve by applying a third-order Savitzky–Golay filter with a window of 301 data points, subtracting the smooth light curve, and clipping out data points more deviant than 1.5× the r.m.s. We performed 10 iterations of this clipping, removing the majority of stellar flares. We then used the exoplanet package (https://github.com/dfm/exoplanet) to simultaneously model the stellar variability and transits. Exoplanet uses several other software packages: Starry for the transit model (https://github.com/rodluger/starry) and celerite (https://github.com/dfm/celerite) for the GP, which we use to model stellar variability. Our GP model consists of two terms; a term to capture long-term trends, and a term to capture the periodic modulation of the star's light curve that is caused by spots on the stellar surface. The latter is a mixture of two stochastically-driven, damped harmonic oscillator terms that can be used to model stellar rotation. It has two modes in Fourier space: one at the rotation period of the star and one at half the rotation period. The transit model is parameterized by two stellar limb-darkening parameters, the log of the orbital period, the log of the stellar density, the time of first transit, the log of the planet-to-star radius ratio, the impact parameter of the transit, orbital eccentricity of the planet, and the periastron angle.

We next run a Markov Chain Monte Carlo (MCMC) to fit for the 9 PLD coefficients (the c_is), a slope + quadratic ramp to represent the rotational modulation of the stellar activity still visible for AU Mic in the Spitzer light curve at 4.5 μm, as well as a transit model including two limb-darkening coefficients for a quadratic limb-darkening law (Extended Data Fig. 2). We leave the photometric uncertainty as a free parameter, which we fit for during the MCMC. Prior to the MCMC, we cut out the dip that occurs during the transit, potentially due to a large spot crossing, from Barycentric Modified Julian Date (BMJD) = 58,524.5 to 58,524.53, to make sure we weren't biasing the transit depth. The systematics-corrected light curve is used in our light-curve modelling in the main text.

Ground-based light-curve analysis

Ref. [21] conducted a dedicated ground-based search for planets transiting AU Mic. One candidate partial transit event ingress was observed (Barycentric Julian Date BJD = 2,453,590.885), with a depth (flux dimming of the star) of ~3%. By itself, this could be attributed to a number of phenomena associated with the star's youth, debris disk, or systematic errors. The photometric precision of this light curve is not sufficient to identify additional transits of AU Mic b or the candidate transit signal from the TESS light curve.

The SuperWASP team monitored AU Mic for seven seasons as part of a larger all-sky survey[22] (Extended Data Fig. 3). We visually inspect the SuperWASP light curve for evidence of any photometry consistent with an ingress or egress from a transiting planet. On several nights, given the ephemeris of AU Mic b, there are photometry data visually similar to an ingress (for example, Julian day (JD) ~2,453,978.40) or an egress (for example, JD ~2,454,232.56). However, the amplitude of the brightness change is comparable to the amplitude of the red (low-frequency) noise in the SuperWASP light curve, and thus these features are probably not real. We do not model or confirm these candidate events, given the stellar activity and relative photometric precision.

The ground-based photometric monitoring[21,22] of AU Mic establishes the long spot lifetimes, which persist for longer than a single observing season as evidenced by the lack of changes in the light curve over many stellar rotations, a defining characteristic of BY Draconis variables. By comparing the TESS, SuperWASP and ref. [21] light curves, it is clear there is spot evolution on a timescale of a few years, as the shape of the phased light curve does differ between the datasets.

Radial-velocity analysis

Seven RV datasets of AU Mic have been obtained by our team or from the literature and archival data, and a detailed analysis to search for additional planets in the AU Mic system is a subject for future work. In this section, we present the utilization of the higher precision radial velocities from iSHELL, HARPS and HIRES (see below) to rule out higher mass companions, correlations with stellar activity, and confirm the planetary nature of AU Mic b by placing an upper limit on its mass. iSHELL[30] is a near-infrared echelle spectrometer with a resolution of $R = 70{,}000$ and a simultaneous grasp of a wavelength range of 300 nm at the 3.0-m NASA Infrared Telescope Facility (IRTF); it is equipped with our custom-built methane isotopologue absorption gas cell for wavelength calibration and instrument characterization[31]. The iSHELL data reduction and RV extraction follows the prescription in ref. [31]. We combine our data with archival observations from the visible wavelength HARPS at the ESO La Silla 3.6-m telescope[32], and the visible wavelength HIRES on the 10-m Keck telescope[33] obtained for the California Planet Survey. All HARPS spectra were extracted and calibrated with the standard ESO Data Reduction Software, and RVs were measured using a least-squares template matching technique[34] (Extended Data Figs. 4–6).

AU Mic is very active relative to a main-sequence dwarf, and we find RV peak-to-peak variations in excess of 400 m s^{-1} in the visible range due to the rotational modulation of stellar activity (r.m.s. = 175 m s^{-1} for HIRES and 115 m s^{-1} for HARPS). With iSHELL, the RVs exhibit stellar activity with a smaller but still substantial peak-to-peak amplitude of ~150 m s^{-1} (r.m.s. = 59 m s^{-1}). Consequently, no individual RV dataset possesses a statistically significant periodogram signal at the period of planet b. This renders the mass detection of a planet with a velocity semi-amplitude smaller than the activity amplitude challenging[35–38].

We perform an MCMC simulation to model the stellar activity with a Gaussian Process (GP) simultaneously with a circular orbit model for AU Mic b using the regression tool RADVEL[39] (Extended Data Fig. 7). Offsets for the velocity zero point of each RV instrument are modelled. We fix the orbital period and time of transit conjunction (orbital phase) for AU Mic b to the best-fit values constrained by the TESS observations. We assume a velocity semi-amplitude prior with a width of 50% of the best-fit value and positive-definite. Owing to the stellar activity and relatively sparse cadence sampling leading to GP model overfitting, no statistically significant constraints on orbital eccentricity are possible; the eccentricity posterior distributions are unconstrained over the range of eccentricities allowed. Thus, for the sake of brevity we present here only scenarios with fixed circular orbits, although eccentric orbits are considered. Constraining the eccentricity (and periastron angle) of AU Mic b will require a more intensive RV cadence and/or new modelling and mitigation of stellar activity beyond a GP model.

The stellar activity is modelled as a GP with a four 'hyper-parameter' auto-correlation function that accounts for the activity amplitude, the rotation period of the star modulating the starspots, and spot lifetimes treated as an autocorrelation decay[37,40]. From photometric time-series, the spot lifetime for AU Mic is observed to be longer than an observing season. Combined with its known rotation period, this enables us to generate priors on the GP hyper-parameters. We use a Jeffrey's prior on the GP hyper-parameter activity amplitudes bounded between 1 and 400 m s^{-1} for the visible, and 1 and 200 m s^{-1} for the near-infrared, a spot decay lifetime prior that is a Gaussian centred on 110 days with a width of 25 days, a stellar rotation period prior of a Gaussian centred on 4.863 days with a width of 0.005 days, and a Gaussian prior centred on 0.388 with a width of 5% for the fourth hyper-parameter. We assess the dependence of our model comparison on the priors and prior widths used for the planet and GP parameters, which yield qualitatively similar results.

We use the MCMC simulations (Extended Data Fig. 8) to compare statistically favoured models obtained from evaluating the model log-likelihoods, AICc (corrected Akaike information criterion) and BIC (Bayesian Information Criterion) statistics (Extended Data Table 1), and to provide robust characterization of model parameter uncertainties (for example, posterior probability distributions). We derive an upper limit to the velocity reflex motion from AU Mic b of $K < 28.9$ m s^{-1} at 3σ confidence, corresponding to a mass upper limit of $M_b < 0.18 M_{Jupiter}$ or $< 3.4 M_{Neptune}$. We restrict our analysis to estimating an upper limit to the mass of AU Mic b for a number of reasons. First, while our statistical analysis favours the detection of AU Mic b, we do not rule out a non-detection at high statistical confidence. Second, our analysis also relies on the assumption that a GP model is an adequate model for stellar activity. Studies of other starspot-dominated convective M dwarfs[38] suggest this is adequate, but additional future observations and modelling efforts are needed, particularly for stars as active as AU Mic. From Kepler photometric time series of main-sequence stars, we demonstrated[40] that stellar activity should not introduce substantial power in densely sampled (approximately nightly) RV time series at orbital periods longer than the stellar rotation period, as is the case for AU Mic b. However, for more sparsely sampled RV cadences such as ours, stellar activity can introduce apparent periodicities at timescales longer than the stellar rotation period that can persist for several seasons[41]. The long-term magnetic activity evolution of AU Mic on timescales >100 days is also neither constrained nor modelled.

Wavelength dependence of stellar activity

At near-infrared wavelengths, the expected stellar activity amplitude depends on the effective temperature contrast of the starspots to the photosphere and the effects of Zeeman broadening[35,42]. If the spot temperature contrast is small (for example, a few hundred kelvin), then the RV (and photometric) amplitude due to the rotational modulation of starspots should scale as $1/\lambda$ to first order. This is the case for the Sun[43]. From the HARPS RV r.m.s., one would expect an RV r.m.s. at 2.3 μm of ~50 m s^{-1} if the HARPS RV r.m.s. is entirely ascribable to stellar activity from cool starspots or plages. However, if the spot temperature contrast is large (for example, >1,000 K), one would expect only a marginal (~10%) reduction in RV stellar activity amplitude in the near-infrared. AU Mic lies close to but slightly above the theoretical expectation for cool starspots with small rather than large spot temperature contrast—showing an RV r.m.s. of 59 m s^{-1}, a reduction of about two-thirds overall in r.m.s. The modelled GP hyper-parameters for the GP amplitudes show a reduction of about one-half from the visible to the near-infrared.

Ref. [21] obtained multi-band photometry of AU Mic over the course of several rotation periods in their search for transiting exoplanets. Ref. [21] demonstrates that AU Mic exhibits a decreased amplitude of photometric variability as a function of wavelength, again consistent with cool starspots with a relatively small temperature contrast (Extended Data Fig. 9). This is also consistent with multi-band photometry of young pre-main-sequence stars and the Sun[44,45].

Host star parameters

We compare the mass derived from transit photometry plus Center for High Angular Resolution Astronomy (CHARA) array radius to pre-main-sequence solar-metallicity isochrones of Baraffe et al.[46]. We logarithmically interpolate onto a finer grid, and fit to the absolute J, H and K_s magnitudes (from 2MASS photometry and the Gaia parallax), the radius derived from CHARA[17] and the Gaia parallax, and the effective temperature. The best-fit ($\chi^2 = 20.7$, $\nu = 3$) age and mass are 19 Myr

and 0.58$M_\odot$; the uncertainties in age and mass are highly correlated, with a 95.4% confidence interval that spans 9–25 Myr, and (0.38–0.63)$M_\odot$.

Future work

Additional RVs are necessary to increase the statistical confidence in the determination and recovery of the orbital parameters for AU Mic b and to search for additional planets. In particular, red-sensitive and near-infrared RVs with a nightly monitoring campaign for at least one season are necessary given the relatively large amplitude and timescale of stellar activity, and if possible to search for additional Neptune-mass and smaller planets. Near-simultaneous chromatic RVs, taken at multiple wavelengths across the visible and near-infrared, and/or polarimetric observations may enable a future analysis that more robustly models the stellar activity than can be accomplished with GP and the non-simultaneous multi-wavelength RVs presented here. Simultaneous multi-wavelength RVs could isolate the chromatic stellar activity signal from the achromatic planet signals. Additionally, AU Mic has a $v\sin i$ value of 8.7 km s^{-1}, and Zeeman Doppler imaging may enable a mapping of the spot configuration on the stellar surface of AU Mic to monitor long-term activity changes.

Future ground- and space-based photometric monitoring, particularly at red and infrared wavelengths, are needed to further constrain the transit parameters. Observing transit timing variations (TTVs) may be possible for this system to search for additional planets, but the analysis will be complicated by the rotational modulation of the starspots and flares. Flares occur frequently during transit, and since AU Mic b potentially crosses active features on the stellar surface, this renders precise transit depth and duration measurements challenging. Here again, simultaneous multi-wavelength photometry could assist in distinguishing the transit signal from stellar activity. In particular, the Spitzer light curve presented here and planned future observations will provide insights into the spot structure of the surface of AU Mic from spot-crossings by AU Mic b for cross-comparison with the Zeeman Doppler imaging maps.

AU Mic b is also an interesting target to search for signatures of its atmosphere, and for extended hydrogen or helium exospheres, with multiple existing and planned near-term instrumentation on the ground and in space. Given its potentially low density, AU Mic b is one of the most favourable targets to search for planetary atmospheres, even taking into account the upper-limit mass measurement. In particular, since the host star AU Mic is a young active star, it may promote the helium mass loss already detected in other Neptune-size bodies[47,48]. Thus, high-dispersion transmission spectroscopy with visible and near-infrared spectrographs, around the 1,083 nm He I and the Hα line, will measure or constrain atmospheric mass loss rate from this young warm planet.

Since the AU Mic system is young, nearby, possesses a debris disk and is a planet that can be observed in transit, it provides an interesting laboratory to explore several theoretical issues. First, simulations should be carried out of the present and past interactions between the inner planet, the possible inner debris disk at <3 AU (ref. [16]), and the outer debris disk including its clumpy structures[7,49,50]. These interactions depend on the masses of both the outer disk and the inner planet, so that this analysis could provide constraints on their properties; moreover, given the 22 Myr age of the star, these integrations can be carried out over the entire possible age of the stellar system. Second, sensitive searches for trace gas could be carried out for this system. Until a few years ago, the classical definition of a debris disk was the secondary generation of dust. Recently, an increasing number of debris disks have shown gas (today up to 17 sources), including the debris disk orbiting β Pic[51], which is rich in carbon, oxygen and nitrogen, perhaps originating from icy grains rich in CO.

Last, it would be useful to compare the properties of AU Mic b with predictions from planet formation/evolution models. If the mass of AU Mic b is close to our upper limit, the observed radius is close to its expected value for a several Gyr-old planet, whereas the predicted contraction timescale of Neptune-size, gas-rich planets is longer than the age of the system[52,53]. These can be reconciled if the planet is substantially less massive than our upper limit. A better mass limit or determination could place interesting constraints on the entropy of planet formation and early thermal evolution.

Data availability

In addition to the figure data available, all raw spectroscopic data are available either in the associated observatory archive or upon request from the corresponding author. The TESS light curve is available at the MAST archive, and the SuperWASP light curve is available at the NASA Exoplanet Archive. Source data are provided with this paper.

Code availability

All code that is not readily available on GitHub is available upon request.

20. Eggen, O. J. Narrow- and broad-band photometry of red stars. II. Dwarfs. *Astrophys. J. Suppl. Ser.* **16**, 49 (1968).
21. Hebb, L. et al. A search for planets transiting the M-dwarf debris disc host, AU Microscopii. *Mon. Not. R. Astron. Soc.* **379**, 63–72 (2007).
22. Pollacco, D. L. et al. The WASP project and the SuperWASP cameras. *Publ. Astron. Soc. Pacif.* **118**, 1407–1418 (2006).
23. Jenkins, J. et al. Overview of the Kepler science processing pipeline. *Astrophys. J.* **713**, L87–L91 (2010).
24. Jenkins, J. et al. The TESS science processing operations center. *Proc. SPIE* **9913**, 99133E (2016).
25. Eastman, J., Gaudi, B. S. & Agol, E. EXOFAST: a fast exoplanetary fitting suite in IDL. *Publ. Astron. Soc. Pacif.* **125**, 83 (2013).
26. Eastman, J. EXOFASTv2: Generalized publication-quality exoplanet modeling code. (ascl:1710.003, Astrophysics Source Code Library, 2017).
27. Kipping, D. Characterizing distant worlds with asterodensity profiling. *Mon. Not. R. Astron. Soc.* **440**, 2164–2184 (2014).
28. Stumpe, M. C. et al. Multiscale systematic error correction via wavelet-based bandsplitting in Kepler data. *Publ. Astron. Soc. Pacif.* **126**, 100 (2014).
29. Smith, J. C. et al. Kepler presearch data conditioning II — a Bayesian approach to systematic error correction. *Publ. Astron. Soc. Pacif.* **124**, 1000 (2012).
30. Rayner, J. T. et al. iSHELL: a construction, assembly and testing. *Proc. SPIE* **9908**, 990884 (2016).
31. Cale, B. et al. Precise radial velocities of cool low mass stars with iSHELL. *Astron. J.* **158**, 170 (2019).
32. Mayor, M. et al. Setting new standards with HARPS. *Messenger* **114**, 20 (2003).
33. Howard, A. W. et al. The California Planet Survey. I. Four new giant exoplanets. *Astrophys. J.* **721**, 1467–1481 (2010).
34. Anglada-Escudé, G. & Butler, R. P. The HARPS-TERRA project. I. Description of the algorithms, performance, and new measurements on a few remarkable stars observed by HARPS. Astrophys. *J. Suppl. Ser.* **200**, 15 (2012).
35. Reiners, A. et al. Detecting planets around very low mass stars with the radial velocity method. *Astrophys. J.* **710**, 432–443 (2010).
36. Tal-Or, L. et al. The CARMENES search for exoplanets around M dwarfs. Radial-velocity variations of active stars in visual-channel spectra. *Astron. Astrophys.* **614**, A122 (2018).
37. Haywood, R. *Hide and Seek: Radial-Velocity Searches for Planets around Active Stars*. PhD thesis, Univ. St Andrews (2015).
38. Barnes, J. R. et al. Recovering planet radial velocity signals in the presence of starspot activity in fully convective stars. *Mon. Not. R. Astron. Soc.* **466**, 1733–1740 (2017).
39. Fulton, B. J., Petigura, E. A., Blunt, S. & Sinukoff, E. RadVel: the radial velocity modeling toolkit. *Publ. Astron. Soc. Pacif.* **130**, 044504 (2018).
40. Vanderburg, A. et al. The Goldilocks trap: stellar activity masquerading as habitable exoplanets. *Mon. Not. R. Astron. Soc.* **459**, 3565 (2016).
41. Nava, C. et al. Exoplanet imitators: a test of stellar activity behavior in radial velocity signals. *Astron. J.* **159**, 23 (2020).
42. Reiners, A. et al. Radial velocity signatures of Zeeman broadening. *Astron. Astrophys.* **552**, A103 (2013).
43. Marchwinski, R. C. et al. Toward understanding stellar radial velocity jitter as a function of wavelength: the Sun as a proxy. *Astrophys. J.* **798**, 63 (2015).
44. Cody, A. M. et al. CSI 2264: Simultaneous optical and infrared light curves of young disk-bearing stars in NGC 2264 with CoRoT and Spitzer — evidence for multiple origins of variability. *Astron. J.* **147**, 82 (2014).
45. Parks, J. R. et al. Periodic and aperiodic variability in the molecular cloud ρ Ophiuchus. Astrophys. *J. Suppl. Ser.* **211**, 3 (2014).
46. Baraffe, I. et al. New evolutionary models for pre-main sequence and main sequence low-mass stars down to the hydrogen-burning limit. *Astron. Astrophys.* **577**, A42 (2015).
47. Nortmann, L. et al. Ground-based detection of an extended helium atmosphere in the Saturn-mass exoplanet WASP-69b. *Science* **362**, 1388–1391 (2018).

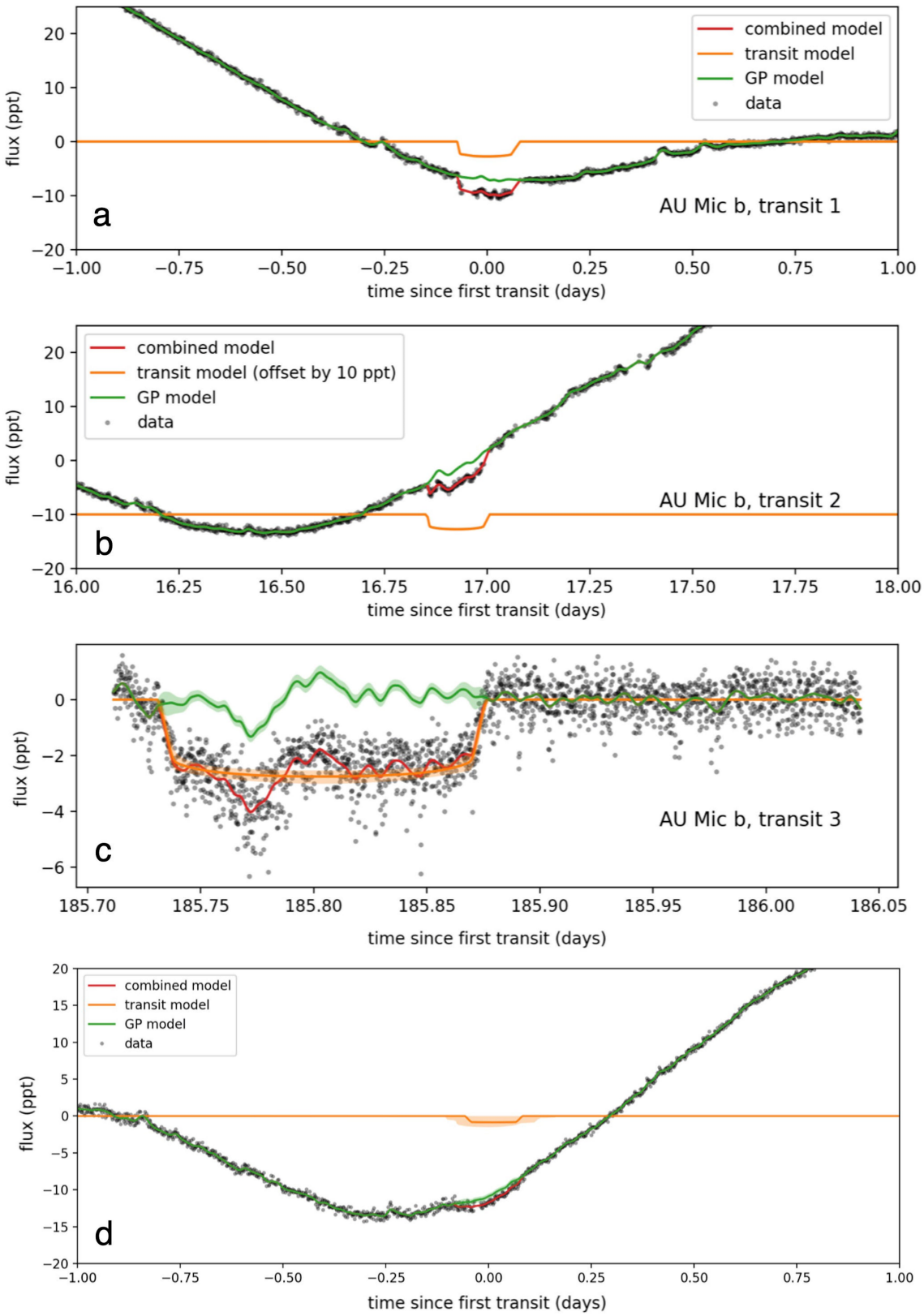
48. Allart, R. et al. Spectrally resolved helium absorption from the extended atmosphere of a warm Neptune-mass exoplanet. *Science* **362**, 1384–1387 (2018).
49. Wang, J. J. et al. Gemini Planet Imager observations of the AU Microscopii debris disk: asymmetries within one arcsecond. *Astrophys. J.* **811**, L19 (2015).
50. Roccatagliata, V. et al. Long-wavelength observations of debris discs around sun-like stars. *Astron. Astrophys.* **497**, 409–421 (2009).
51. Wilson, P. A. et al. Detection of nitrogen gas in the β Pictoris circumstellar disc. *Astron. Astrophys.* **621**, A121 (2019).
52. Nettelmann, N. et al. Uranus evolution models with simple thermal boundary layers. *Icarus* **275**, 107 (2016).
53. Linder, E. et al. Evolutionary models of cold and low-mass planets: cooling curves, magnitudes, and detectability. *Astron. Astrophys.* **623**, A85 (2019).

Acknowledgements This work was supported by grants to P.P. from NASA (award 16-APROBES16-0020 and support from the Exoplanet Exploration Program) and the National Science Foundation (Astronomy and Astrophysics grant 1716202), the Mount Cuba Astronomical Foundation and George Mason University start-up funds.The NASA Infrared Telescope Facility is operated by the University of Hawaii under contract NNH14CK55B with NASA. Funding for the TESS mission is provided by NASA's Science Mission directorate. Some of the data presented here were obtained at the W. M. Keck Observatory, which is operated as a scientific partnership among the California Institute of Technology, the University of California and NASA. The Observatory was made possible by the generous financial support of the W. M. Keck Foundation. The authors wish to recognize and acknowledge the very significant cultural role and reverence that the summit of Maunakea has always had within the indigenous Hawaiian community. We are most fortunate to have the opportunity to conduct observations from this mountain. This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with NASA under the Exoplanet Exploration Program. Some of the data presented in this paper were obtained from the Mikulski Archive for Space Telescopes (MAST). The Space Telescope Science Institute is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. This research has made use of the services of the ESO Science Archive Facility, based on observations collected at the European Organisation for Astronomical Research in the Southern Hemisphere with the HARPS spectrometer. This work has made use of data from the European Space Agency (ESA) mission Gaia, processed by the Gaia Data Processing and Analysis Consortium (DPAC). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement. MINERVA-Australis is supported by Australian Research Council LIEF Grant LE160100001, Discovery Grant DP180100972, Mount Cuba Astronomical Foundation, and institutional partners University of Southern Queensland, MIT, Nanjing University, George Mason University, University of Louisville, University of California Riverside, University of Florida and University of Texas at Austin. This work was partly supported by JSPS KAKENHI grant numbers JP18H01265 and 18H05439, JST PRESTO grant number JPMJPR1775, NSFC grant number 11673011 and MINECO grant ESP2016-80435-C2-2-R. D.D. acknowledges support for this work provided by NASA through Hubble Fellowship grant HST-HF2-51372.001-A awarded by the Space Telescope Science Institute. B.P.B. acknowledges support from National Science Foundation grant AST-1909209. J.W. and P.G. acknowledge support from the Heising-Simons Foundation 51 Pegasi b fellowship.

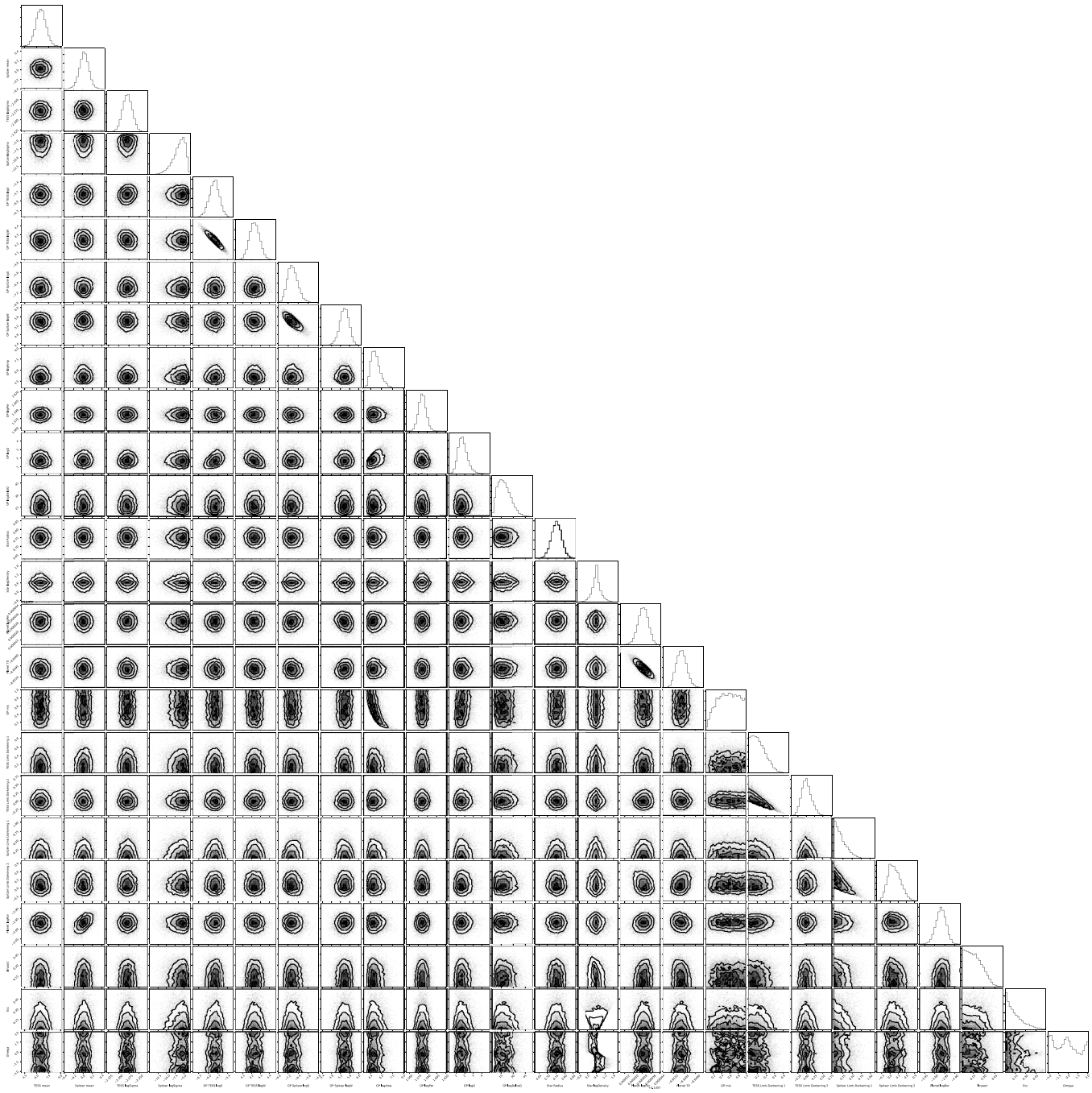
Author contributions P.P.: lead author, principal investigator for CSHELL/iSHELL gas cell and observations, analysis and interpretation. J.G., P.G., B.C., A.T., S.X.W., R.W.: CSHELL/iSHELL data reduction and forward model codes. W.M.: RADVEL analysis. T.B., D.D., S.Q., D.F.-M., E. Gilbert, C. Huang, D.K., E.K., E.V.Q., A.V.: analysis of TESS light curve. K.S., K.C., N.N., E.P., J.P.: follow-up ground-based observations. I.J.M.C., D.A.B., P.L., E.N.: Spitzer light curve. D.F., B.T., C. Hellier: inspection of ground-based light curves. D.W.L.: TRES. G.A.-E.: HARPS. G.R., R.V., S.S., J.N.W., J.M.J.: TESS mission architects. S.R., A.K., S.D., J.T.: TESS mission. F.A., M.C., M.K., A.R., V.R., J.W.: disk physics. D.A.A., J.E.S., A.Y.: flare analysis. C. Beichman, M.B., C. Brinkworth, D.R.C., S.R.K., B.M., S.M.M., K.v.B.: CSHELL/iSHELL instrumentation. B.P.B., C.J.B., J.T.C., J. Horner, J.K., J.O., C.G.T., R.A.W., D.J.W., H.Z.: MINERVA-Australis. A.C., C.D., E.F., C.G., F.G., R.H., T.H., J.H., C.K., N.L., M.M., T.M., A.N., J.T., B.W., D.W., P.Z.: CSHELL/iSHELL observers. E.J.G.: stellar parameters. A.W.H.: Keck HIRES.

Competing interests The authors declare no competing interests.

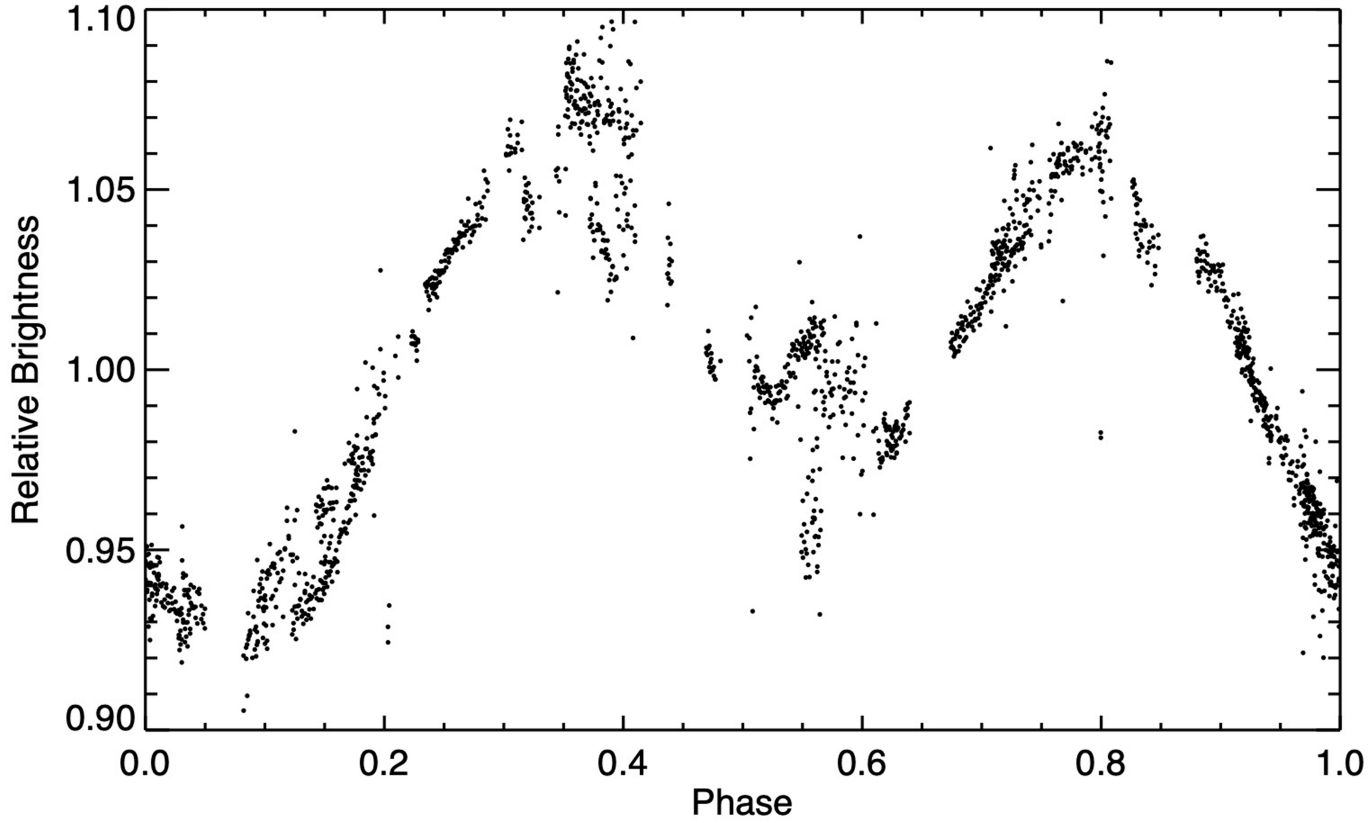
Additional information
Correspondence and requests for materials should be addressed to P.P.
Peer review information *Nature* thanks Suzanne Aigrain and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



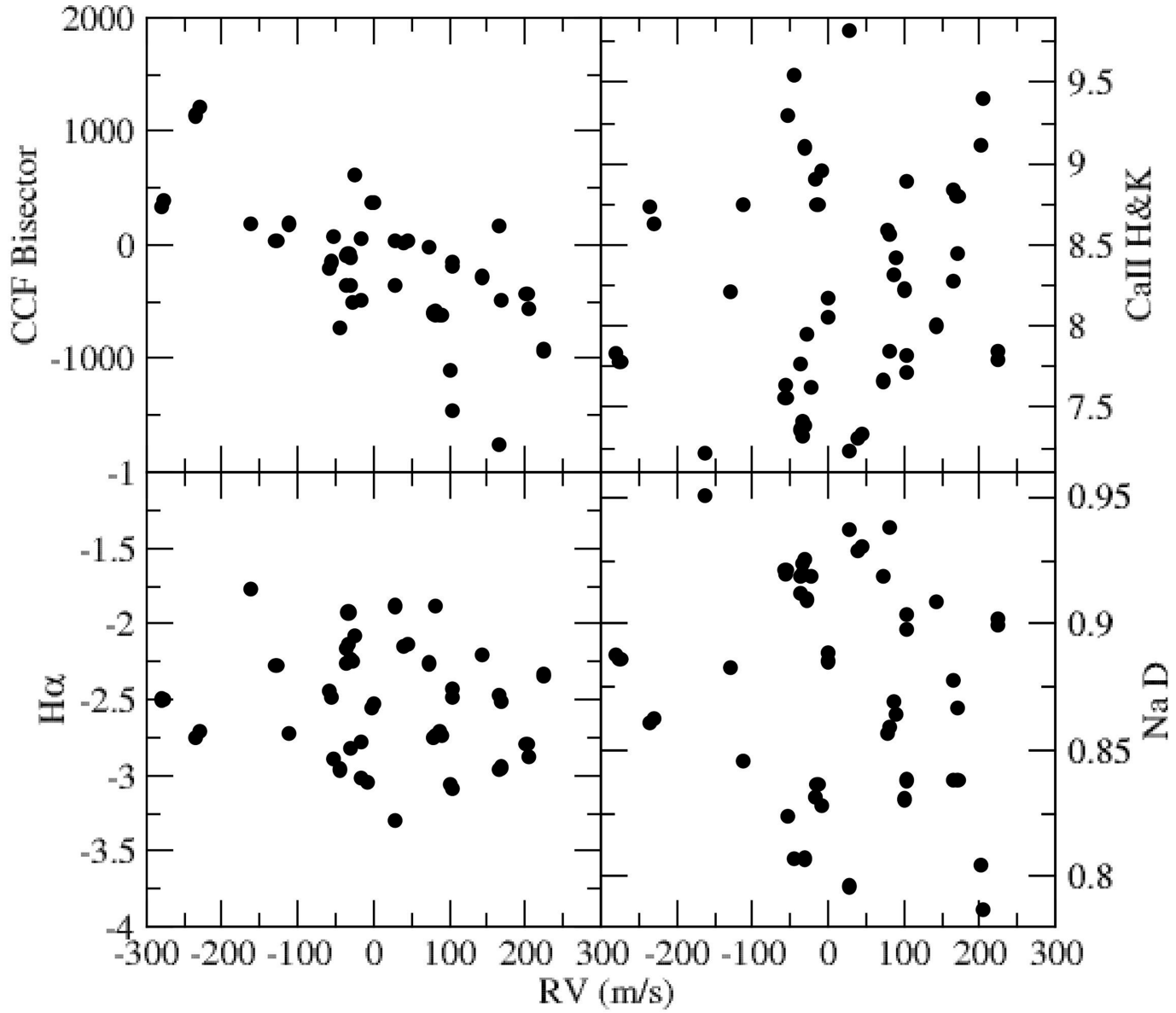
Extended Data Fig. 1 | TESS and Spitzer light curves for AU Mic centred on four transit events. a, **b**, Two TESS transits (respectively 1 and 2) for AU Mic b, with the model components plotted as indicated in the key. A flare is present during the egress of the first transit of AU Mic b, and a flare is present just after the ingress during the second transit of AU Mic b. Although this is unfortunate timing, flares of this amplitude are pervasive throughout the TESS light curve for AU Mic, and complicate the recovery of these events from automated transit search algorithms. **c**, The Spitzer transit observation of AU Mic b. The deviations in transit are not instrumental and will be the subject of a future paper, and are likely to be related to the planet crossing large active regions on the stellar surface (key from **a** and **b** applies here). **d**, The ~1 p.p.t. candidate single transit event seen in the TESS light curve. For all panels, 1σ measurement uncertainties are suppressed for visual clarity and are <1 p.p.t. 1σ model uncertainties in transit are shown as shaded regions.



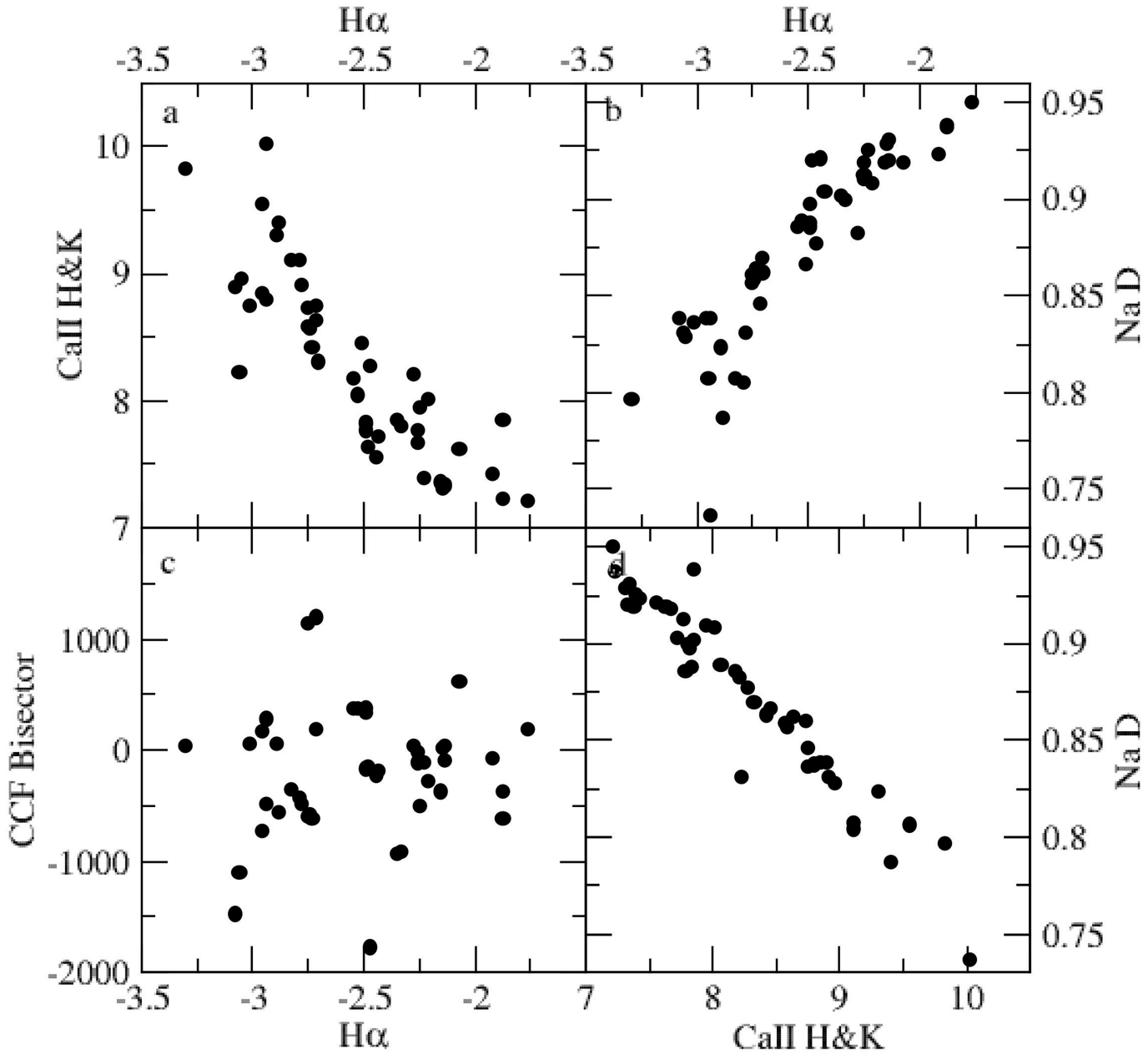
Extended Data Fig. 2 | MCMC corner plot for custom combined Spitzer and TESS light-curve analysis for AU Mic. The full set of model parameters are shown, with the posterior probability distributions along the diagonal, the others are the two-dimensional parameter covariance plots.



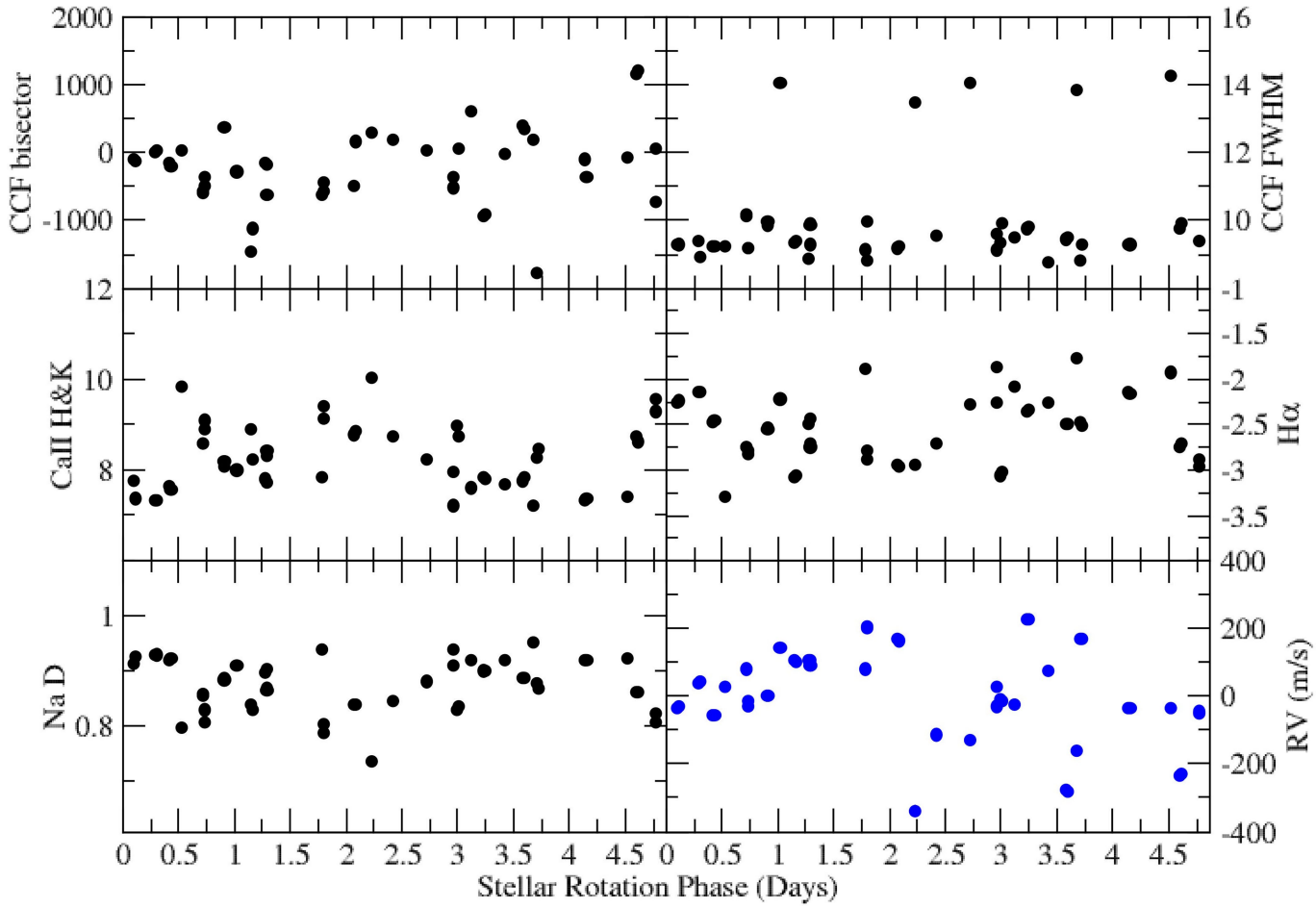
Extended Data Fig. 3 | One season (July to October 2007) of SuperWASP light curves for AU Mic from the NASA Exoplanet Archive, phase-folded to the rotation period of the star. Measurements with large photometric uncertainties (>5%) have been excluded from the plot. 1σ measurement uncertainties are suppressed for visual clarity and are typically <1% but occasionally up to 5% at phases where there is more apparent vertical scatter in the measurement values themselves.



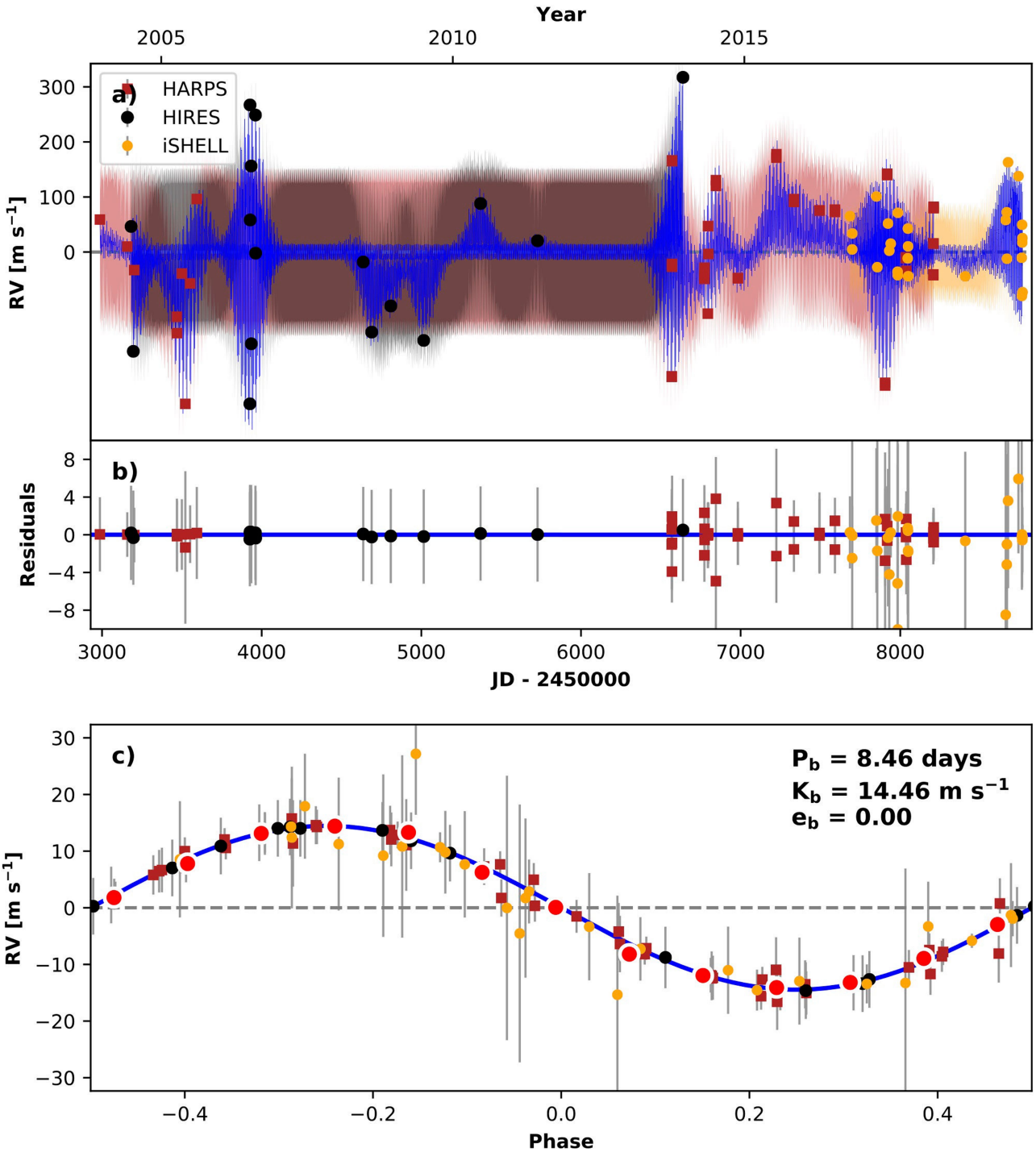
Extended Data Fig. 4 | Correlation plots of the standard HARPS stellar activity indicators with the RVs. The bisector values for the cross-correlation function ('CCF bisector'), but not the activity indicators (Hα, Na D, Ca II H and K), show a correlation with the RVs, with substantial remaining scatter. Formal uncertainties are smaller than the plotted symbols.



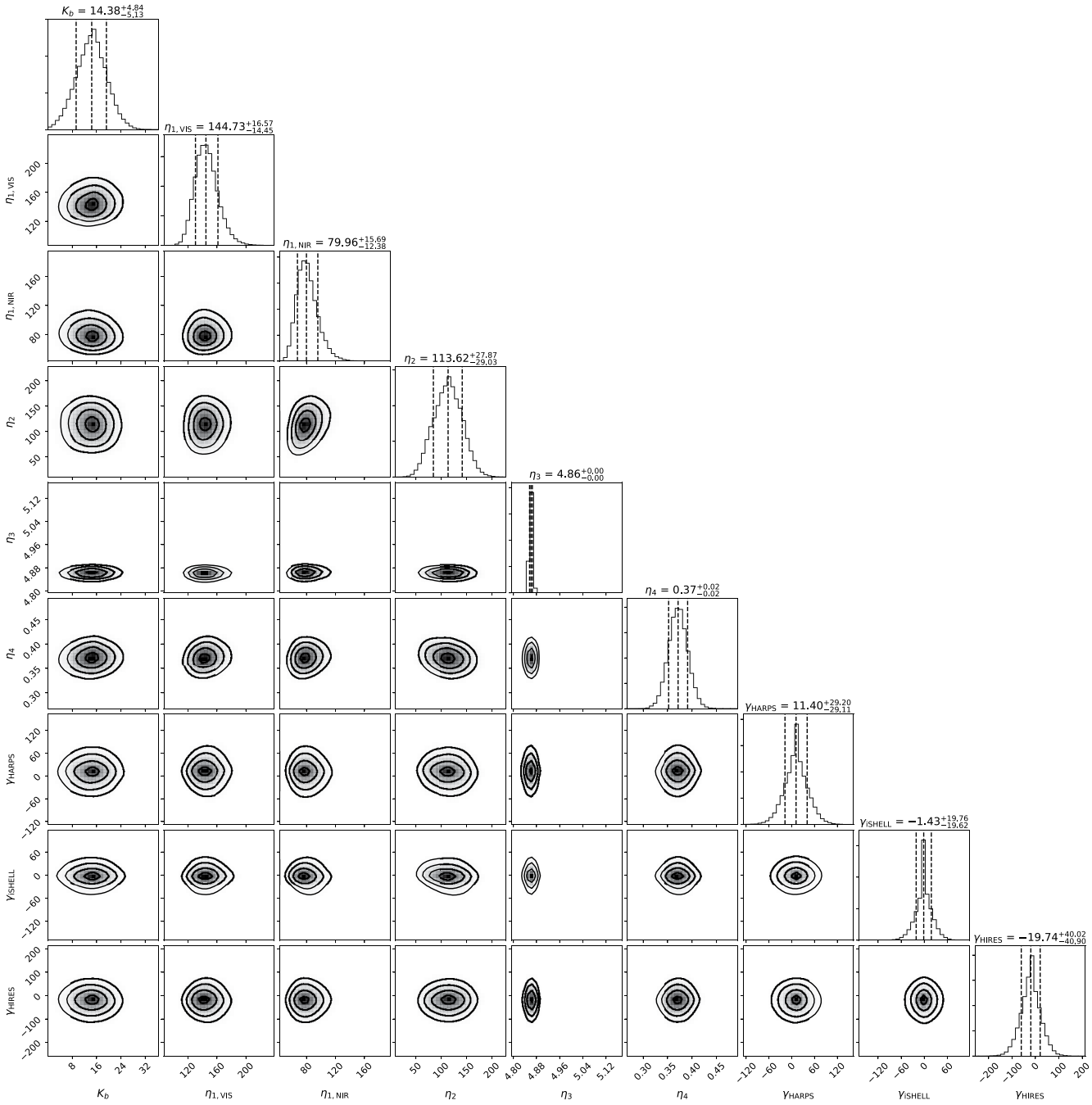
Extended Data Fig. 5 | Correlation plots of the HARPS activity indicators with each other. The activity indicators Ca II H and K, Hα, and Na D are strongly correlated with one another, but not with the RVs or with the CCF bisector.



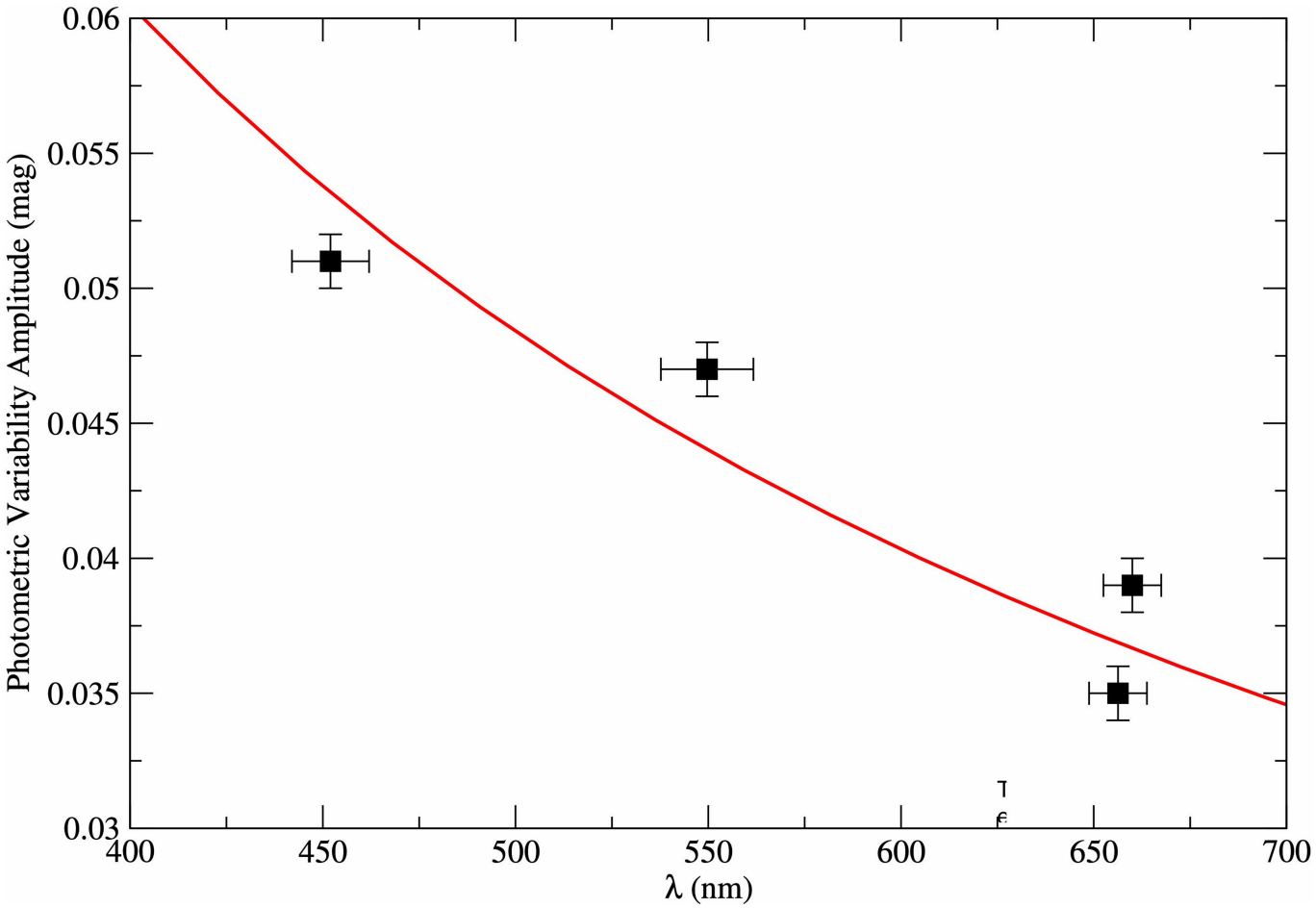
Extended Data Fig. 6 | The HARPS RVs and standard activity indicators, phase folded to the rotation period of the star. Blue circles, HARPS RVs; black circles, standard activity indicators. None of the activity indicators show a statistically significant trend with the period of AU Mic b. The Ca and Na activity indicators appear to show (by eye) some cyclic variation with the rotation period of the star. Formal uncertainties are smaller than the plotted symbols.



Extended Data Fig. 7 | RV time-series of AU Mic, with fitting residuals, and phased to the orbital period of AU Mic b. Shown are data from three spectrometers: iSHELL (yellow circles), HIRES (black circles) and HARPS (red squares). Uncertainties shown are 1σ for HARPS and iSHELL. For HIRES, a 5 m s^{-1} minimum 1σ uncertainty is adopted, although the formal 1σ uncertainties are smaller for all but one epoch at 5.43 m s^{-1}. The maximum-likelihood best fit model is overlaid in blue, with shaded regions indicating the 1σ model confidence interval, with a separate GP for each dataset indicated with different coloured shaded regions. **b**, Model-subtracted residuals, with the same colours as in **a**. Because our RVs are undersampled with respect to the stellar rotation period[38], the GP best-fit model overfits the AU Mic RV time-series. **c**, RV measurements are phased to the orbital period of AU Mic b, and binned in phase (red circles). The blue curve is a maximum-likelihood best-fit circular orbit model, after subtracting the best fit GP model of stellar activity and the modelled instrument offsets. The plot is labelled with the best-fit orbital period P_b, velocity semi-amplitude K_b, and the assumed circular orbit ($e_b = 0$).



Extended Data Fig. 8 | RADVEL MCMC corner plot for the model parameters for the iSHELL, HARPS and HIRES RV datasets. Along the diagonal are the one-dimensional posterior probability distributions for a given model parameter; the others are the two-dimensional parameter covariance plots.



Extended Data Fig. 9 | Photometric variability amplitudes obtained contemporaneously in four different bandpasses. The amplitudes (black squares) are from ref. [21]. The horizontal error bars correspond to the effective bandpass widths, and the 1σ vertical error bars are set to 1 mmag. A 1/λ trend is shown in red, as would be expected for cool starspots with relatively small temperature contrast[35].

Extended Data Table 1 | Model comparison results

AU Mic Model (all include GP & data set offsets)	RV data sets	Free Parameters	Number of RV epochs	Best-fit model rms	log-likelihood	BIC	AICc	Δ AICc between favoured model	AICc qualitative comparison
b	iSHELL, HARPS, HIRES	9	91	2.68	-505.14	1050.88	1030.50	0	Favoured Model
Gaussian Process only	iSHELL, HARPS, HIRES	8	91	3.02	-509.05	1054.18	1035.85	5.35	Strongly disfavoured

Article

Entanglement-based secure quantum cryptography over 1,120 kilometres

https://doi.org/10.1038/s41586-020-2401-y

Received: 15 July 2019

Accepted: 13 May 2020

Published online: 15 June 2020

Check for updates

Juan Yin[1,2,3], Yu-Huai Li[1,2,3], Sheng-Kai Liao[1,2,3], Meng Yang[1,2,3], Yuan Cao[1,2,3], Liang Zhang[2,3,4], Ji-Gang Ren[1,2,3], Wen-Qi Cai[1,2,3], Wei-Yue Liu[1,2,3], Shuang-Lin Li[1,2,3], Rong Shu[2,3,4], Yong-Mei Huang[5], Lei Deng[6], Li Li[1,2,3], Qiang Zhang[1,2,3], Nai-Le Liu[1,2,3], Yu-Ao Chen[1,2,3], Chao-Yang Lu[1,2,3], Xiang-Bin Wang[2], Feihu Xu[1,2,3], Jian-Yu Wang[2,3,4], Cheng-Zhi Peng[1,2,3]✉, Artur K. Ekert[7,8] & Jian-Wei Pan[1,2,3]✉

Quantum key distribution (QKD)[1–3] is a theoretically secure way of sharing secret keys between remote users. It has been demonstrated in a laboratory over a coiled optical fibre up to 404 kilometres long[4–7]. In the field, point-to-point QKD has been achieved from a satellite to a ground station up to 1,200 kilometres away[8–10]. However, real-world QKD-based cryptography targets physically separated users on the Earth, for which the maximum distance has been about 100 kilometres[11,12]. The use of trusted relays can extend these distances from across a typical metropolitan area[13–16] to intercity[17] and even intercontinental distances[18]. However, relays pose security risks, which can be avoided by using entanglement-based QKD, which has inherent source-independent security[19,20]. Long-distance entanglement distribution can be realized using quantum repeaters[21], but the related technology is still immature for practical implementations[22]. The obvious alternative for extending the range of quantum communication without compromising its security is satellite-based QKD, but so far satellite-based entanglement distribution has not been efficient[23] enough to support QKD. Here we demonstrate entanglement-based QKD between two ground stations separated by 1,120 kilometres at a finite secret-key rate of 0.12 bits per second, without the need for trusted relays. Entangled photon pairs were distributed via two bidirectional downlinks from the Micius satellite to two ground observatories in Delingha and Nanshan in China. The development of a high-efficiency telescope and follow-up optics crucially improved the link efficiency. The generated keys are secure for realistic devices, because our ground receivers were carefully designed to guarantee fair sampling and immunity to all known side channels[24,25]. Our method not only increases the secure distance on the ground tenfold but also increases the practical security of QKD to an unprecedented level.

Our experimental arrangement is shown in Fig. 1. The two receiving ground stations are located at Delingha (37°22′44.43″N, 97°43′37.01″E; altitude 3,153 m) in Qinghai province, and Nanshan (43°28′31.66″N, 87°10′36.07″E; altitude 2,028 m) in Xinjiang province, China. The physical distance between Delingha and Nanshan is 1,120 km. To optimize the receiving efficiencies, both the two ground telescopes are newly built with a diameter of 1.2 m, specifically designed for the entanglement distribution experiments. All the optical elements, such as mirrors, in the telescopes maintain polarization.

The satellite is equipped with a compact spaceborne entangled photon source with a weight of 23.8 kg. A periodically poled $KTiOPO_4$ crystal inside a Sagnac interferometer is pumped in both the clockwise and anticlockwise directions simultaneously by a continuous-wave laser with a wavelength centred at 405 nm and a linewidth of 160 MHz, and generates down-converted polarization-entangled photon pairs at 810 nm close to the form of $|\Psi\rangle_{12} = (|H\rangle_1|V\rangle_2 + |V\rangle_1|H\rangle_2)/\sqrt{2}$, where $|H\rangle$ and $|V\rangle$ denote the horizontal and vertical polarization states, respectively, and the subscripts 1 and 2 denote the two output spatial modes. The entangled photon pairs are then collected and guided by two single-mode fibres to two independent transmitters equipped in the satellite. Both transmitters have a near-diffraction-limited far-field divergence of about 10 μrad. Under a pump power of 30 mW, the source distributes up to 5.9×10^6 entangled photon pairs per second.

The photons are collected by the telescopes on two optical ground stations. For each one, the follow-up optics is installed on one of the rotating arms and rotates along with the telescope. As shown in Fig. 1c,

[1]Hefei National Laboratory for Physical Sciences at the Microscale and Department of Modern Physics, University of Science and Technology of China, Hefei, China. [2]Shanghai Branch, CAS Center for Excellence in Quantum Information and Quantum Physics, University of Science and Technology of China, Shanghai, China. [3]Shanghai Research Center for Quantum Science, Shanghai, China. [4]Key Laboratory of Space Active Opto-Electronic Technology, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai, China. [5]The Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China. [6]Shanghai Engineering Center for Microsatellites, Shanghai, China. [7]Mathematical Institute, University of Oxford, Oxford, UK. [8]Centre for Quantum Technologies, National University of Singapore, Singapore, Singapore. ✉e-mail: pcz@ustc.edu.cn; pan@ustc.edu.cn

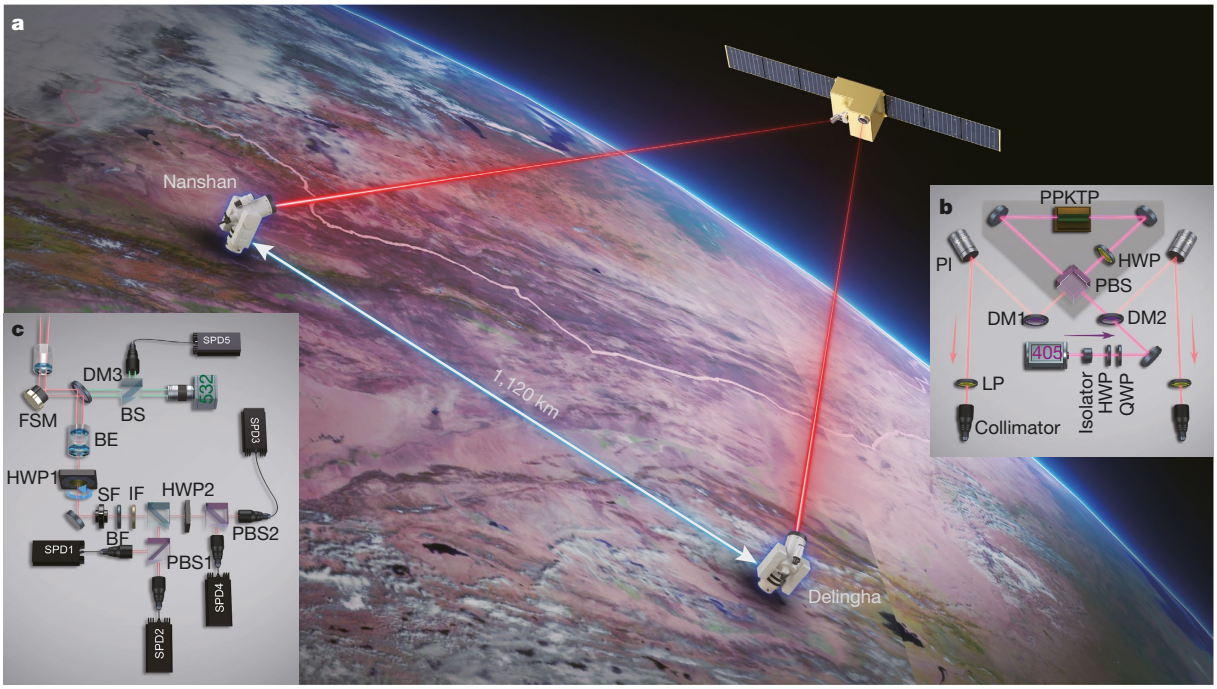


Fig. 1 | Overview of the experimental set-up of entanglement based quantum key distribution. a, An illustration of the Micius satellite and the two ground stations. Image credit: Fengyun-3C/Visible and Infrared Radiometer, with permission (2020). The satellite flies in a Sun-synchronous orbit at an altitude of 500 km. The physical distance between Nanshan and Delingha ground station is 1,120 km. **b**, The spaceborne entangled-photon source. A free space isolator is used to minimize back reflection to the 405-nm pump laser. A pair of off-axis concave mirrors is used to focus the pump laser and collimate the down-converted photon pairs. PBS, polarization beam splitter; DM, dichroic mirror; LP, long-pass edge filter; PI, piezo steering mirror; HWP, half-wave plate; QWP, quarter-wave plate; PPKTP, periodically poled $KTiOPO_4$. **c**, The follow-up optic at the optical ground station. The tracking and synchronization laser is separated from the signal photon by DM3 and detected by the single photon detector (SPD5). The spatial filter (SF), broad-bandwidth filter (BF) and interference filter (IF) are used to filter out the input light in frequency and spatial domains. BS, beam splitter; BE, beam expander; FSM, fast steering mirror.

a beam splitter, a half-wave plate and two polarized beam splitters are combined to analyse the polarization of the entangled photons randomly in the bases of $Z \in \{|H\rangle, |V\rangle\}$ and $X \in \{|+\rangle, |-\rangle\}$, where $|\pm\rangle = (|H\rangle \pm |V\rangle)/\sqrt{2}$. After being transmitted or reflected by the beam splitter and polarized beam splitters, the photons are collected by four multimode fibres with the core diameter of 105 μm and detected by four single photon detectors (SPDs) respectively. We carefully selected the four SPDs to ensure that the detector efficiency is better than 53%, the efficiency consistency is better than 98.5% and the dark counts are less than 100 counts per second (see Extended Data Table 1 for details). A motorized half-wave plate (HWP1) is used to compensate the relative rotation between the transmitter and the receiver, where the correction angle offsets are calculated in advance. The entangled photons are filtered in both the frequency and spatial domains to satisfy the fair sampling assumption and to guarantee practical security. In particular, an extra field diaphragm, consisting of two lenses with focal length of 8 mm and a pinhole of 100 μm, is used as the spatial filter to unify the field of view of different channels, where the field of view is narrowed to 27 μrad. A broad-bandwidth filter and a narrow-bandwidth filter of 5 nm are used to reject frequency side channels. These frequency filters can also help to reduce the background counts. The output signals of the SPDs are recorded by a time-to-digital converter.

To optimize the link efficiency, we develop cascaded multistage acquiring, pointing and tracking systems both in the satellite transmitters and the optical ground station receivers, achieving a tracking accuracy of 2 μrad and 0.4 μrad, respectively. The beacon laser (532 nm, 10 kHz) from the satellite is also used as a synchronization laser. It is sampled, frontier identified and recorded by the same time-to-digital converter as well as quantum signals. The distant time-to-digital converters are first roughly synchronized using a global positioning system (GPS) one-pulse-per-second (1PPS) signal. As the frequency of the synchronization laser is relatively stable, a least-squares method is used to fit the selected pulses, which can eliminate the time jitter of synchronization detectors. The time synchronization accuracy of entangled photon pairs is 0.77 ns (1σ). We set a narrow coincidence time gate of 2.5 ns to reduce the accidentally coincident events.

The satellite flies along a Sun-synchronous orbit, and comes into both Delingha's and Nanshan's view once every night, starting at around 2:00AM Beijing time and lasting for a duration of 285 s (>13° elevation angle for both ground stations). Figure 2a plots the physical distances from the satellite to Delingha and Nanshan during one orbit, together with the sum channel length of the two downlinks. As shown in Fig. 2b, the measured overall two-downlink channel attenuation varies from 56 dB to 71 dB. As compared to previous experiment[23], this two-photon count rate, and thus the signal-to-noise ratio, is greatly improved. To increase the collection efficiency for downlink entangled photons, we have upgraded both the main system of the telescope and the follow-up optics. For the main system, we improved the receiving efficiency by recoating the main lens (+1.5 dB) and redesigning the high-efficiency beam expander (+0.9 dB). For the follow-up optics, we increased the collection efficiency through optical pattern matching, especially shortening the optical path by 20 cm to avoid beam spreading by 0.65 mm (+0.6 dB).

As a result, we have increased the collection efficiency of each satellite-to-ground link by a factor of about 2 over the previous experiment[23]. This was quantified by measuring the single-downlink efficiencies of each ground station for several orbits. The best-orbit data were taken on a clear night with no clouds in the sky and no haze near the ground, which had the highest atmospheric transmittance (Extended Data Fig. 1). Under these conditions, the link efficiency is related only

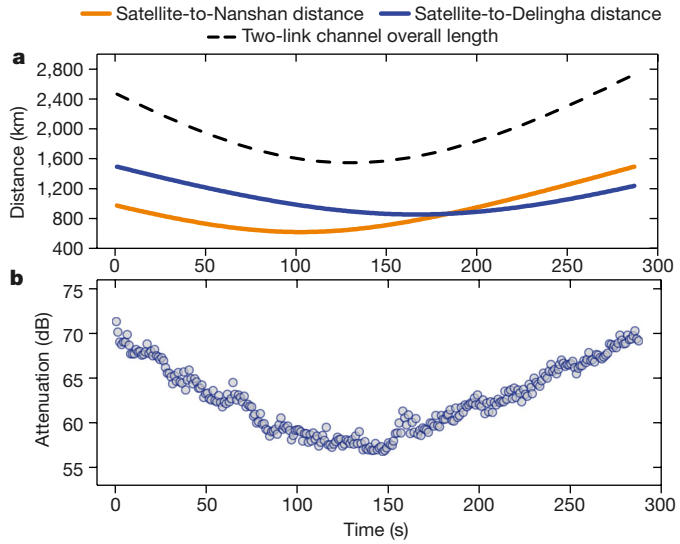


Fig. 2 | Distances and attenuations from satellite to Nanshan (Delingha). **a**, A typical two-downlink trial from satellite to Nanshan, and to Delingha, lasts about 285 s (>13° elevation angle for both ground stations) in a single pass of the satellite. The distance from satellite to Nanshan (Delingha) is from 618 km (853 km) to about 1,500 km, and the total length of the two downlinks varies from 1,545 km to 2,730 km. **b**, The measured satellite-to-ground two-downlink channel attenuation.

to the distance between the satellite and the ground (Extended Data Fig. 2). These data were selected to calibrate the improvement of the link efficiency, and a 3-dB enhancement in the collection efficiency was observed for each satellite-to-ground link (Extended Data Fig. 3). Overall, the collection efficiency of the two-photon distribution was improved by a factor of about 4 over the previous experiment[23].

To realize secure QKD against side-channel attacks, we add several single-mode filters to the receiver, which slightly decreases the collection efficiency. Even so, the system efficiency (with filters) still improves by a factor of about 2. A comparison of the results of this work and the previous experiment[23] is shown in Extended Data Table 2. We observe an average two-photon count rate of 2.2 Hz, with a signal-to-noise ratio of 15:1. The sifted key rate for QKD is 1.1 Hz. This enhancement is remarkable, because it decreases the quantum bit error rate (QBER) from about 8.1% (ref. [23]) to about 4.5%, thus enabling the realization of satellite-based entanglement QKD (Extended Data Fig. 4).

The entanglement-based QKD system was carefully designed to provide practical security against physical side channels[21,22]. We note that entanglement-based QKD is naturally source-independent[16,17], which guarantees that the system is secure against loopholes in the source. All we need is to ensure the security on the detection sides, that is, the two optical ground stations. In general, the side channels on the detection side primarily violate the key assumption of fair sampling. To guarantee this assumption, we add a series of filters with different degrees of freedom, including frequency, spatial and temporal modes, and implement countermeasures for the correct operation of the single-photon detectors.

Specifically, great attention has been paid to detection attacks, including: detector-related attack[26–28], wavelength-dependent attack[29], spatial-mode attack[30], and other possible side-channels. We have implemented countermeasures to all the above known attacks (see Methods and Extended Data Table 3). For the side channels targeting the operation of detectors, such as blinding attack[26], we install additional monitoring circuits. In particular, we install an additional circuit to monitor the anode of the load resistance in the detection circuit to counter the blinding attack (Extended Data Fig. 5). If there is a bright laser pulse illumination, the output of the monitoring circuit will exceed a secure threshold voltage and trigger the alarm (Fig. 3b). For the time-shift attack[27] and the dead-time attack[28], our countermeasure is to operate the detector in free-running mode, in which the detector records all the detection events and post-selects the detection windows such that the detection efficiency is guaranteed to be at a nominal level. For the side channels in other optical domains (Fig. 1c), we use optical filters to filter out the input light and eliminate the mismatch in the frequency and spatial domains. In particular, we use two cascaded broad-bandwidth and narrow-bandwidth filters (Fig. 3a) to eliminate the frequency dependency[29] of the transmission/reflection ratio of the beam splitter (Extended Data Fig. 6). Spatial filters are added to ensure identical efficiencies for different detectors (Fig. 3c), thus eliminating the spatially dependent loopholes[30]. Consequently, the secret key, generated by our QKD system, is practically secure for realistic devices.

To verify the entanglement established between the two distant optical ground stations, we use the distributed entangled photons for the Bell test with the Clauser–Horne–Shimony–Holt (CHSH)-type inequality[31], which is given by

$$S = |E(\varphi_1, \varphi_2) - E(\varphi_1, \varphi_2') + E(\varphi_1', \varphi_2) + E(\varphi_1', \varphi_2')| \le 2$$

where E is the joint correlation with measurement angles of the Delingha optical ground station and the Nanshan optical ground station, respectively. The angles are randomly selected from (0, π/8), (0, 3π/8), (π/4, π/8) and (π/4, 3π/8) to close the locality loophole. We run 1,021 trials of the Bell test during an effective time of 226 s. The observed result for parameter S is 2.56 ± 0.07, with a violation of the CHSH–Bell inequality $S < 2$ by 8 standard deviations (see Extended Data Table 4 for details). The Bell violation provides evidence of high-quality entanglement between the entangled photons observed over 1,120 km apart.

In our entanglement-based QKD demonstration, we adopted the BBM92 protocol[3], in which the measurements by Alice and Bob are symmetric, that is, each of them requires two measurement bases, that is, the Z (H/V) basis and the X (+/−) basis. As mentioned above, using filtering and monitoring, we guarantee that the single-photon detections were conducted on a nearly two-dimensional subspace and the system detection efficiencies for the four polarization states could be well characterized to satisfy the fair sampling condition without Eve's tampering. Experimentally, we have characterized the system detection efficiency of each detection path, where the efficiency mismatch has an upper bound of 1.47%. This efficiency mismatch is considered in the privacy amplification (PA) of the post-processing of the secret key rate (see Methods). Moreover, we use the post-processing to handle double clicks, by randomly assigning a classical bit, as well as the dead-time effect, by removing the sequential detections after a click. These implementations can ensure that the secret keys produced are secure against the issues of known side channels.

Following the security analysis for an uncharacterized source[19], the asymptotic secret key rate R_Z for the post-processed bits in the Z basis is given by:

$$R_Z \ge Q_Z[1 - f_e H(E_Z) - H(E_X)]$$

where Q_Z is the sifted key where Alice and Bob select the Z basis, f_e is the error correction inefficiency, and E_Z and E_X are the QBER in the Z and X bases, respectively. The analysis for the X basis is the same. The total asymptotic secret key rate is $R_A = R_Z + R_X$. The detailed security analysis for the finite key rate R_F, which takes into account the finite key size[32,33] and the detection efficiency mismatch, is shown in Methods.

Experimentally, we obtained 6,208 initial coincidences within 3,100 s of data collection. Discarding the events for which the two optical ground stations had chosen different bases, we obtained 3,100 bits of sifted key with 140 erroneous bits, which corresponded to an averaged QBER of 4.51% ± 0.37%. The QBERs in the H/V and +/− bases (Z and X bases) are, respectively, 4.63% ± 0.51% and 4.38% ± 0.54%. For

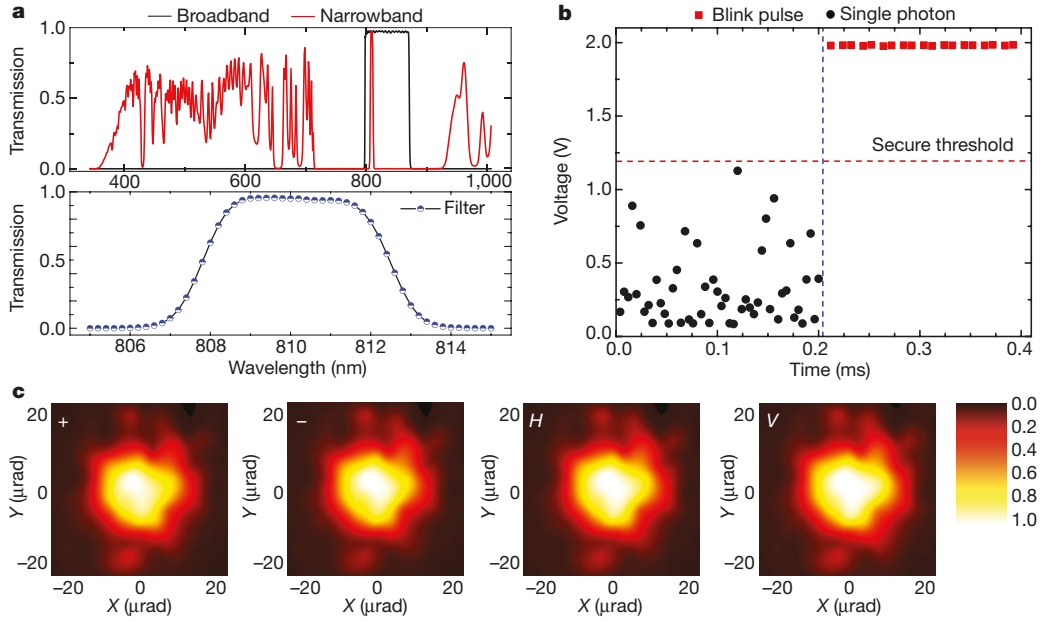


Fig. 3 | Monitoring and filtering against side channels. a, The transmission of broad-bandwidth and narrow-bandwidth wavelength filters. **b**, The output of monitoring circuit with/without blinding attack. Without blinding attack, the outputs are random avalanching single-photon-detection signals (black dots). With blinding attack (starting from 0.20 ms), the output signals are at around 2 V, which is clearly above the security threshold, thus triggering the security alarm. **c**, The system detection efficiency of the four polarizations in the spatial domain. With the spatial filter, the four efficiencies are identical. The colour scale shows the measured efficiencies normalized to the maximum efficiency.

the sifted bits, we performed an error correction with Hamming code and achieved an error correction inefficiency of $f_e = 1.19$. After the error correction and the PA, we obtained a secure key rate of $R_A = 0.43$ bits per second in the asymptotic limit of the infinitely long key. With a failure probability $\varepsilon = 10^{-10}$, the finite key rate is $R_F = 0.12$ bits per second (see Table 1 for a summary). In total, we obtained a 372-bit secret key. Compared to directly transmitting the entangled photons over a distance of 1,120 km using commercial ultralow-loss optical fibres (with a loss of 0.16 dB km^{-1}), we estimate that the effective link efficiency, and thus the secret key rate, of the satellite-based method is eleven orders of magnitude higher. The secure distance substantially outperforms previous entanglement-based QKD experiments[12,34].

In summary, we have demonstrated entanglement-based QKD between two ground stations separated by 1,120 km. We increase the link efficiency of the two-photon distribution by a factor of about 4 compared to the previous work[23] and obtain a finite-key secret key rate of 0.12 bits per second. The brightness of our spaceborne entangled photon source can be increased by about two orders of magnitude in our latest research[35], which could readily increase the average final key to tens of bits per second or tens of kilobits per orbit. The entanglement-based quantum communication could be combined with quantum repeaters[21] for general quantum communication protocols and distributed quantum computing[36]. Hence, our work paves the way towards entanglement-based global quantum networks. Overall, the results increase the secure distance of practical QKD on the ground from 100 km to more than 1,000 km without the need for trusted relays, thus representing an important step towards a truly robust and unbreakable cryptographic method for remote users over arbitrarily long distances.

Table 1 | Experimental results of entanglement-based QKD over 1,120 km

Parameter	Q	E_Z	E_X	R_A	R_F
Value	1.00 bps	4.63% ± 0.51%	4.38% ± 0.54%	0.43 bps	0.12 bps

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2401-y.

1. Bennett, C. H. & Brassard, G. Quantum cryptography: public key distribution and coin tossing. In *Proc. Int. Conf. on Computers, Systems and Signal Processing* 175–179 (1984).
2. Ekert, A. K. Quantum cryptography based on Bell's theorem. *Phys. Rev. Lett.* **67**, 661 (1991).
3. Bennett, C. H., Brassard, G. & Mermin, N. D. Quantum cryptography without Bell's theorem. *Phys. Rev. Lett.* **68**, 557 (1992).
4. Peng, C.-Z. et al. Experimental long-distance decoy-state quantum key distribution based on polarization encoding. *Phys. Rev. Lett.* **98**, 010505 (2007).
5. Rosenberg, D. et al. Long-distance decoy-state quantum key distribution in optical fiber. *Phys. Rev. Lett.* **98**, 010503 (2007).
6. Yin, H.-L. et al. Measurement-device-independent quantum key distribution over a 404 km optical fiber. *Phys. Rev. Lett.* **117**, 190501 (2016).
7. Boaron, A. et al. Secure quantum key distribution over 421 km of optical fiber. *Phys. Rev. Lett.* **121**, 190502 (2018).
8. Liao, S.-K. et al. Satellite-to-ground quantum key distribution. *Nature* **549**, 43 (2017).
9. Liao, S.-K. et al. Space-to-ground quantum key distribution using a small-sized payload on Tiangong-2 Space Lab. *Chin. Phys. Lett.* **34**, 090302 (2017).
10. Yin, J. et al. Satellite-to-ground entanglement-based quantum key distribution. *Phys. Rev. Lett.* **119**, 200501 (2017).
11. Schmitt-Manderbach, T. et al. Experimental demonstration of free-space decoy-state quantum key distribution over 144 km. *Phys. Rev. Lett.* **98**, 010504 (2007).
12. Ursin, R. et al. Entanglement-based quantum communication over 144 km. *Nat. Phys.* **3**, 481 (2007).
13. Elliott, C. et al. Current status of the DARPA quantum network. In *Quantum Information and Computation III* Vol. 5815, 138–150 (International Society for Optics and Photonics, 2005).
14. Peev, M. et al. The SECOQC quantum key distribution network in Vienna. *New J. Phys.* **11**, 075001 (2009).
15. Chen, T.-Y. et al. Field test of a practical secure communication network with decoy-state quantum cryptography. *Opt. Express* **17**, 6540 (2009).
16. Sasaki, M. et al. Field test of quantum key distribution in the Tokyo QKD network. *Opt. Express* **19**, 10387–10409 (2011).
17. Qiu, J. et al. Quantum communications leap out of the lab. *Nature* **508**, 441 (2014).
18. Liao, S.-K. et al. Satellite-relayed intercontinental quantum network. *Phys. Rev. Lett.* **120**, 030501 (2018).
19. Koashi, M. & Preskill, J. Secure quantum key distribution with an uncharacterized source. *Phys. Rev. Lett.* **90**, 057902 (2003).

20. Ma, X., Fung, C.-H. F. & Lo, H.-K. Quantum key distribution with entangled photon sources. *Phys. Rev. A* **76**, 012307 (2007).
21. Briegel, H.-J., Dur, W., Cirac, J. I. & Zoller, P. Quantum repeaters: the role of imperfect local operations in quantum communication. *Phys. Rev. Lett.* **81**, 5932–5935 (1998).
22. Yang, S.-J., Wang, X.-J., Bao, X.-H. & Pan, J.-W. An efficient quantum light–matter interface with sub-second lifetime. *Nat. Photon.* **10**, 381 (2016).
23. Yin, J. et al. Satellite-based entanglement distribution over 1200 kilometers. *Science* **356**, 1140 (2017).
24. Lo, H.-K., Curty, M. & Tamaki, K. Secure quantum key distribution. *Nat. Photon.* **8**, 595 (2014).
25. Xu, F., Ma, X., Zhang, Q., Lo, H.-K. & Pan, J.-W. Secure quantum key distribution with realistic devices. *Rev. Mod. Phys.* **92**, 025002 (2020).
26. Lydersen, L. et al. Hacking commercial quantum cryptography systems by tailored bright illumination. *Nat. Photon.* **4**, 686 (2010).
27. Zhao, Y., Fung, C.-H., Qi, B., Chen, C. & Lo, H.-K. Quantum hacking: experimental demonstration of time-shift attack against practical quantum-key-distribution systems. *Phys. Rev. A* **78**, 042333 (2008).
28. Weier, H. et al. Quantum eavesdropping without interception: an attack exploiting the dead time of single-photon detectors. *New J. Phys.* **13**, 073024 (2011).
29. Li, H.-W. et al. Attacking a practical quantum-key-distribution system with wavelength-dependent beam-splitter and multiwavelength sources. *Phys. Rev. A* **84**, 062308 (2011).
30. Sajeed, S. et al. Security loophole in free-space quantum key distribution due to spatial-mode detector-efficiency mismatch. *Phys. Rev. A* **91**, 062301 (2015).
31. Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880 (1969).
32. Koashi, M. Simple security proof of quantum key distribution based on complementarity. *New J. Phys.* **11**, 045018 (2009).
33. Tomamichel, M., Lim, C. C. W., Gisin, N. & Renner, R. Tight finite-key analysis for quantum cryptography. *Nat. Commun.* **3**, 634 (2012).
34. Peng, C.-Z. et al. Experimental free-space distribution of entangled photon pairs over 13 km: towards satellite-based global quantum communication. *Phys. Rev. Lett.* **94**, 150501 (2005).
35. Cao, Y. et al. Bell test over extremely high-loss channels: towards distributing entangled photon pairs between earth and the moon. *Phys. Rev. Lett.* **120**, 140405 (2018).
36. Ladd, T. D. et al. Quantum computers. *Nature* **464**, 45–53 (2010).

Methods

Implementation against device imperfections

In practice, the imperfections of realistic QKD implementations may introduce deviations (or side channels) from the idealized models used in the security analysis. Eve might exploit these imperfections and launch quantum attacks[24]. Our entanglement-based QKD implementation is designed and characterized to provide practical security against both known quantum attacks and potential future loopholes.

The entanglement-based QKD is naturally source-independent[2,19]. All we need is to consider the side channels properly at the detection stage. Here, we design a detection system, choosing apparatus under strict criteria for satisfying the underlying security assumptions, and performing careful characterizations to test those assumptions. We note that our implementation is based on trusted and characterized devices, that is, in a device-dependent scenario. The implementations are mostly common techniques, but we can maintain immunity to all known detection attacks, including: detector efficiency-mismatch attack[37], time-shift attack[27,38], detector-blinding attack[26,39], detector-damage attack[40], detector dead-time attack[28], wavelength-dependent attack[29], spatial-mode attack[30], and other possible side channels[24]. In Extended Data Table 3, we list the reported attacks against the detection, as well as our countermeasures to avert them. In the following, we will give a more detailed description.

Efficiency-mismatch attack. In practice, it is difficult to manufacture two SPDs with the same responses for different degrees of freedom. That is, practical SPDs present efficiency mismatch. With the efficiency mismatch, Eve can partially control which detector clicks by subtly sending desired signals to Bob[37]. For example, most of QKD systems use two gated avalanche photodiode detectors, which produce a time-dependent efficiency mismatch. Eve can perform a time-shift attack[27,38], by shifting the arrival time of each signal, so that Bob's detection results are biased depending on the time shift. Our strategy to counter the time-shift attack is that our detector works in free-running mode. We record all the detection events and post-select the detection windows such that the detection efficiency is guaranteed to be at a nominal level. For efficiency mismatch in other degrees of freedom[37], we use optical filters to filter out the input light and eliminate the mismatch in the frequency and spatial modes.

Detector-blinding attack. In the detector-blinding attack[26], Eve uses a continuous bright laser illumination to force SPDs to work in the linear mode. The SPDs are then no longer sensitive to single photons, and are converted into classical intensity detectors. Eve can control which detector clicks by sending Bob properly tailored classical pulses. In the laser damage attack[40], Eve can use a strong damaging laser illumination to change the properties of the SPDs completely. To counter the detector-blinding attack and the laser-damage attack, as illustrated in Extended Data Fig. 5, we install an additional circuit to monitor the anode of the load resistance in the detection circuit. We test the attack during the experiment by sending a bright laser pulse illumination. These results are shown in Fig. 3b. In normal operation (without blinding pulses), the output voltage of the monitoring circuit is below 1.2 V, corresponding to standard avalanching signals. At time $t \approx 0.2$ ms, Eve performs the blinding attack using 12 μW and a 2-μs-long laser pulse at a repetition rate of 100 kHz. The output of the monitoring circuit clearly exceeds 1.2 V, because a large current caused by the bright laser illumination passes through the load resistance. Consequently, we could set a secure threshold on the voltage of monitoring circuit: if the voltage is higher than the threshold, it exposes the blinding attack.

Detector dead-time attack. The basic principle of this attack is the dead-time effect of a SPD[28]. After a detection event, a detector does not respond to the incoming photons during a time window ranging from several nanoseconds to tens of microseconds. If Bob has a detection event during a time period when one detector is in the dead-time period, while the other one is active, Eve could easily infer which detector has a click. Our detector works in the free-running mode, and all detection events are collected. The countermeasure is that we monitor the status of the detectors and use only those detection events for which all detectors are active to generate keys.

Beam-splitter attack. In a polarization-based QKD system, Bob typically exploits an 1×2 beam splitter to passively choose the measurement basis. In the standard case, a photon will randomly pass through the beam splitter, thus randomly selecting a rectilinear basis or a diagonal basis. However, in practice, the splitting ratio of the beam splitter is wavelength-dependent, that is, the centre wavelength has a coupling ratio of 50:50, whereas the coupling ratio varies for other wavelengths. Consequently, Eve can control the measurement basis by sending Bob photons with different wavelength[29]. To avoid this attack, we use broad-bandwidth and narrow-bandwidth wavelength filters to filter the input light on Bob's station. The characterizations of these two filters are shown in Fig. 3a. The beam splitter ratio within the filtered bandwidth is characterized in Extended Data Fig. 6.

Spatial-mode attack. In a free-space QKD system, the detector has different sensitivities for different spatial-mode photons, especially when the detector is coupled with a multi-mode fibre. Eve could exploit the spatial-mode efficiency mismatch and perform the spatial-mode attack[30]. To counter this attack, we place a spatial filter in front of the beam splitter to make the efficiencies of different detection paths uniform. With the spatial filter, the characterization of the detection efficiency in spatial domain is shown in Fig. 3c.

In general, the practical security of implementation is essentially guaranteed by the fair-sampling assumption. The countermeasures to the abovementioned attacks comprise the use of active components to guarantee the fair-sampling assumption. In the frequency mode, broad-band and narrow-band frequency filters are employed to filtering the input light. In the temporal mode, free-running detectors are applied to post-select the time windows of detection events. In the spatial mode, spatial filters are placed before the collimating lens of measurement devices. In polarization mode, we use the polarization encoding for QKD, thus monitoring the QBER to ensure the security. In future, we may also combine our entanglement-based QKD system with the measurement-device-independent QKD protocol[41] to make detection immune to all detector attacks.

Security analysis

The main goal of our security analysis is to calculate the practical security rate by considering the issues of the finite-key size and device imperfections. We remark that our security analysis is for entanglement-based QKD with trusted and characterized devices, that is, in a device-dependent scenario[42]. We start with a security proof for an ideal QKD protocol by following the Shor–Preskill security proof[43]. We then extend the security analysis to the practical case of the finite-key effect by using the approach of uncertainty relation for smooth entropies[33]. Finally, we extend the analysis to address the security issues of device imperfections by using the Gottesman–Lo–Lütkenhaus–Preskill (GLLP) framework[44].

Ideal QKD refers to the case where an infinite number of signals are generated and the devices to run the QKD protocol are as perfect as described by theoretical models. The security proof for ideal QKD was established in the early 2000s by Mayers[45], Lo and Chau[46] and Shor and Preskill[43].

Shor and Preskill employed the idea of the Calderbank–Shor–Steane quantum error correcting code to provide a simple framework for security proof. In an entanglement-based QKD such as the BBM92 protocol[3], when Alice and Bob both measure quantum signals in the Z

basis, an error may occur when the outcomes are different. We can call it a bit error. The phase error can be defined as the hypothetical error if those quantum signals were measured in the basis complementary to the Z basis. In the Shor–Preskill security proof, the bit error correction is classical error correction and the phase error correction is PA. The crucial part is to perform the PA, in which one needs to estimate the phase error rate. For the key bits measured in the Z basis, the phase error rate can be estimated by measuring the key bits in the X basis. The Z-basis security rate for ideal QKD is given by

$$R_Z \geq Q_Z[1 - H(E_Z) - H(E_X)]$$

where Q_Z is the sifted key rate per signal in which both Alice and Bob select the Z basis, E_Z and E_X are the QBER in the Z and X bases, and $H(\chi) = -\chi\log_2\chi - (1-\chi)\log_2(1-\chi)$. Similarly, secret keys can also be generated in the X basis, and the analysis for the rate R_X is the same. The total ideal key rate is $R_A = R_Z + R_X$. Note that an entangled source is basis-independent (or uncharacterized), and the security proof for QKD with an uncharacterized source is given in ref. [19].

We remark that in order for a successful estimation of PA, one needs to make sure the sampling in the complementary basis is fair, which in practical realizations raises two major issues: the finite-key effect (that is, statistical fluctuations) and device imperfections (that is, violating the fair sampling), discussed below.

Finite-key analysis

We first define the security in the finite-key scenario with the composable security definition framework[47,48]. A secure key should satisfy two requirements. First, the key bit strings possessed by Alice and Bob need to be identical, that is, to be correct. Second, from the view of anyone other than Alice and Bob, say Eve, the key bit string should be uniformly distributed, that is, should be secret. Practical issues, such as the finite data size and non-ideal error correction, mean that Alice and Bob cannot generate an ideal key via QKD. In reality, it is reasonable to allow the key to have small failure probabilities, ε_{cor} and ε_{sec}, for correctness and secrecy. We say that the QKD protocol is ε-secure with $\varepsilon \geq \varepsilon_{cor} + \varepsilon_{sec}$, if it is ε_{cor}-correct and ε_{sec}-secret[48]. Specifically, we define k_a and k_b to be the key bit strings obtained by Alice and Bob. A QKD protocol is defined to be ε_{cor}-correct if the probability satisfies $\Pr(k_a \neq k_b) \leq \varepsilon_{cor}$. A QKD protocol is defined in trace distance to be ε_{sec}-secret, if $[(1-P_{abort})/2]\|\rho_{AE} - U_A \otimes \rho_E\| \leq \varepsilon_{sec}$, where ρ_{AE} is the classical quantum state describing the joint state of k_a and Eve's system ρ_E, U_A is the uniform mixture of all possible values of k_a, and P_{abort} is the probability that the protocol aborts.

There are two main approaches to analyse the finite-key security of QKD: one is based on smooth min/max entropy[33,48] and the other one is based on complementarity[32]. Recently, these two approaches have been proved to be unified[49]. The estimation of the phase error rate is the most important part of the Shor–Preskill security analysis. Owing to statistical fluctuations in the finite-key case, the phase error rate used for evaluating the amount of PA cannot be measured accurately. Instead, Alice and Bob can bound the phase error rate via certain complementary measurements[32,33]. Specifically, for the Z-basis security key in entanglement-based QKD, Alice and Bob can bound the underlying phase error rate E_X' by sampling the qubits in the X basis. This is a typical random sampling problem. We can use the Serfling inequality[50] to estimate the probability that the average error on the sample deviates from the average error on the total string[51]. We obtain the upper bound for E_X' as

$$E_X' \leq E_X + \sqrt{\frac{(n_X+1)\log(1/\varepsilon_{sec})}{2n_X(n_X+n_Z)}}$$

where n_Z and n_X are the number of coincident counts in the Z and X bases.

By using the approach of the uncertainty relation for smooth entropies[33], the Z-basis secret key length l_Z is given by

$$l_Z = n_Z - n_Z H\left[E_X + \sqrt{\frac{(n_Z+1)\log\left(\frac{1}{\varepsilon_{sec}}\right)}{2n_X(n_X+n_Z)}}\right] - f_e n_Z H(E_Z) - \log\frac{2}{\varepsilon_{cor}\varepsilon_{sec}^2}.$$

Similarly, the X-basis finite-key secret key length l_X can be calculated, and the total key length is $l = l_Z + l_X$.

Security proof for imperfect devices

In practice, owing to device imperfections, there exist deviations between realistic QKD systems and the ideal QKD protocol[24]. To achieve practical security in a QKD system, Alice and Bob need to characterize these imperfections carefully and take them into account in the practical security analysis. Notably, a general framework for security analysis with realistic devices was established in ref. [44]. In this framework, Alice and Bob need to characterize their devices to see how much deviation there is from the ideal ones assumed in the security proofs. One can employ typical distance measures, like fidelity and trace distance, to quantify the deviation, and then consider this deviation in PA.

Our entanglement-based QKD is source-independent, which ensures that the imperfections in the source can be ignored. All we need is to carefully characterize the imperfections in the detection side. In general, the (known and to be known) side channels on the detection side[26–30,38–40] primarily violate the key assumption of fair sampling. We perform implementations by following the squashing model[44] to guarantee the fair sampling assumption. In a squashing model, an arbitrary quantum state (from the channel) is first projected to a two-dimensional subspace before the Z and X measurements. So, we implement a series of single-mode filters in different degrees of freedom, including the frequency, spatial and temporal modes. Nonetheless, practical filters normally have finite bandwidth, which will cause small deviations for detection efficiencies, that is, a detection efficiency mismatch[52,53]. Our security proof for imperfect devices will primarily consider the deviation of the detection efficiency, and analyse this imperfection into the PA by following the GLLP framework[44].

We assume the lower bound of detection efficiency is η_0, so the detection efficiency of the ith detector can be written as $\eta_0(1+\delta_i)$, where δ_i quantifies the deviation of efficiency. Suppose that if we can add attenuation with transmittance $1/(1+\delta_i)$ just before the ith detector, then we would obtain equal efficiency for all detectors. In doing so, the number of Z-bits (or X-bits) will be reduced by a fraction, upper bounded by $\Delta = 1 - 1/(1+\delta)^2$. In our experiment, we quantify that δ_i is upper bounded by $\delta_i \leq 1.47\%$ (see Extended Data Table 1). This deviation can be considered in PA, that is, the estimation of phase error rate as $E_X'/(1-\Delta)$ (ref. [44]). Overall, after considering the finite-key size effect and the efficiency deviation, the secret key length L_Z is given by:

$$L_Z = n_Z - n_Z H\left[\frac{E_X + \sqrt{\frac{(n_Z+1)\log\left(\frac{1}{\varepsilon_{sec}}\right)}{2n_X(n_X+n_Z)}}}{1-\Delta}\right] - f_e n_Z H(E_Z) - n_Z\Delta - \log\frac{2}{\varepsilon_{cor}\varepsilon_{sec}^2}.$$

The analysis of the secret key length L_X for the key bits in the X basis is the same. The total finite-key length is $L = L_Z + L_X$.

Data availability

The data that support the findings of this study are available from the corresponding authors on reasonable request.

37. Makarov, V., Anisimov, A. & Skaar, J. Effects of detector efficiency mismatch on security of quantum cryptosystems. *Phys. Rev. A* **74**, 022313 (2006).
38. Qi, B., Fung, C.-H.F., Lo, H.-K. & Ma, X. Time-shift attack in practical quantum cryptosystems. *Quantum Inf. Comput*. **7**, 73 (2007).
39. Gerhardt, I. et al. Experimentally faking the violation of Bell's inequalities. *Phys. Rev. Lett*. **107**, 170404 (2011).
40. Bugge, A. N. et al. Laser damage helps the eavesdropper in quantum cryptography. *Phys. Rev. Lett*. **112**, 070503 (2014).
41. Lo, H.-K., Curty, M. & Qi, B. Measurement-device-independent quantum key distribution. *Phys. Rev. Lett*. **108**, 130503 (2012).
42. Scarani, V. et al. The security of practical quantum key distribution. *Rev. Mod. Phys*. **81**, 1301–1350 (2009).
43. Shor, P. & Preskill, J. Simple proof of security of the BB84 quantum key distribution protocol. *Phys. Rev. Lett*. **85**, 441 (2000).
44. Gottesman, D., Lo, H.-K., Lütkenhaus, N. & Preskill, J. Security of quantum key distribution with imperfect devices. *Quantum Inf. Comput*. **4**, 325 (2004).
45. Mayers, D. J. Unconditional security in quantum cryptography. *J. Assoc. Comput. Mach*. **48**, 351–406 (2001).
46. Lo, H. K. & Chau, H. F. Unconditional security of quantum key distribution over arbitrarily long distances. *Science* **283**, 2050 (1999).
47. Ben-Or, M., Horodecki, M., Leung, D. W., Mayers, D. & Oppenheim, J. In *Proc. 2nd Int. Conf. on Theory of Cryptography (TCC'05)* 386–406 (Springer, 2005).
48. Renner, R. *Security of quantum key distribution. PhD thesis*, ETH Zurich (2005); preprint at https://arxiv.org/abs/quant-ph/0512258.
49. Tsurumaru, T. Leftover hashing from quantum error correction: unifying the two approaches to the security proof of quantum key distribution. Preprint at https://arxiv.org/abs/1809.05479 (2018).
50. Serfling, R. J. Probability inequalities for the sum in sampling without replacement. *Ann. Stat*. **2**, 39–48 (1974).
51. Curty, M. et al. Finite-key analysis for measurement-device-independent quantum key distribution. *Nat. Commun*. **5**, 3732 (2014).
52. Fung, C.-H. F., Tamaki, K., Qi, B., Lo, H.-K. & Ma, X. Security proof of quantum key distribution with detection efficiency mismatch. *Quantum Inf. Comput*. **9**, 131–165 (2009).
53. Marøy, Ø., Lydersen, L. & Skaar, J. Security of quantum key distribution with arbitrary individual imperfections. *Phys. Rev. A* **82**, 032337 (2010).

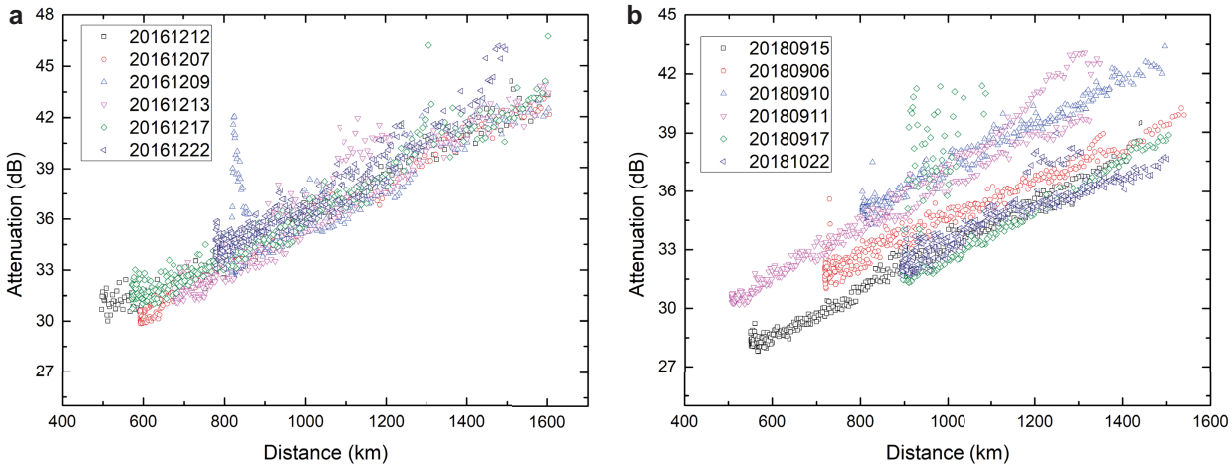
Acknowledgements We acknowledge discussions with X. Ma and C. Jiang. We thank colleagues at the National Space Science Center, China Xi'an Satellite Control Center, National Astronomical Observatories, Xinjiang Astronomical Observatory, Purple Mountain Observatory, and Qinghai Station for their management and coordination. We thank G.-B. Li, L.-L. Ma, Z. Wang, Y. Jiang, H.-B. Li, S.-J. Xu, Y.-Y. Yin, W.-C. Sun and Y. Wang for their long-term assistance in observation. This work was supported by the National Key R&D Program of China (grant number 2017YFA0303900), the Shanghai Municipal Science and Technology Major Project (grant number 2019SHZDZX01), the Anhui Initiative in Quantum Information Technologies, Science and Technological Fund of Anhui Province for Outstanding Youth (grant number 1808085J18) and the National Natural Science Foundation of China (grant numbers U1738201, 61625503, 11822409, 11674309, 11654005 and 61771443).

Author contributions C.-Z.P., A.K.E. and J.-W.P. conceived the research. J.Y., C.-Z.P. and J.-W.P. designed the experiments. J.Y., Y.-H.L., S.-K.L., M.Y., Y.C., J.-G.R., S.-L.L., C.-Z.P. and J.-W.P. developed the follow-up optics and monitoring circuit. J.Y., Y.-M.H., C.-Z.P. and J.-W.P. developed the efficiency telescopes. J.Y., S.-K.L., Y.C., L.Z., W.-Q.C., R.S., L.D., J.-Y.W., C.-Z.P. and J.-W.P. designed and developed the satellite and payloads. J.Y., L.Z., W.-Q.C., W.-Y.L. and C.-Z.P. developed the software. F.X., X.-B.W., A.K.E. and J.-W.P. performed the security proof and analysis. L.L., Q.Z., N.-L.L., Y.-A.C., X.-B.W., F.X., C.-Z.P., A.K.E. and J.-W.P. contributed to the theoretical study and implementation against device imperfections. F.X., C.-Y.L., C.-Z.P. and J.-W.P. analysed the data and wrote the manuscript, with input from J.Y., Y.-H.L., M.Y., Y.C. and A.K.E. All authors contributed to the data collection, discussed the results and reviewed the manuscript. J.-W.P. supervised the whole project.

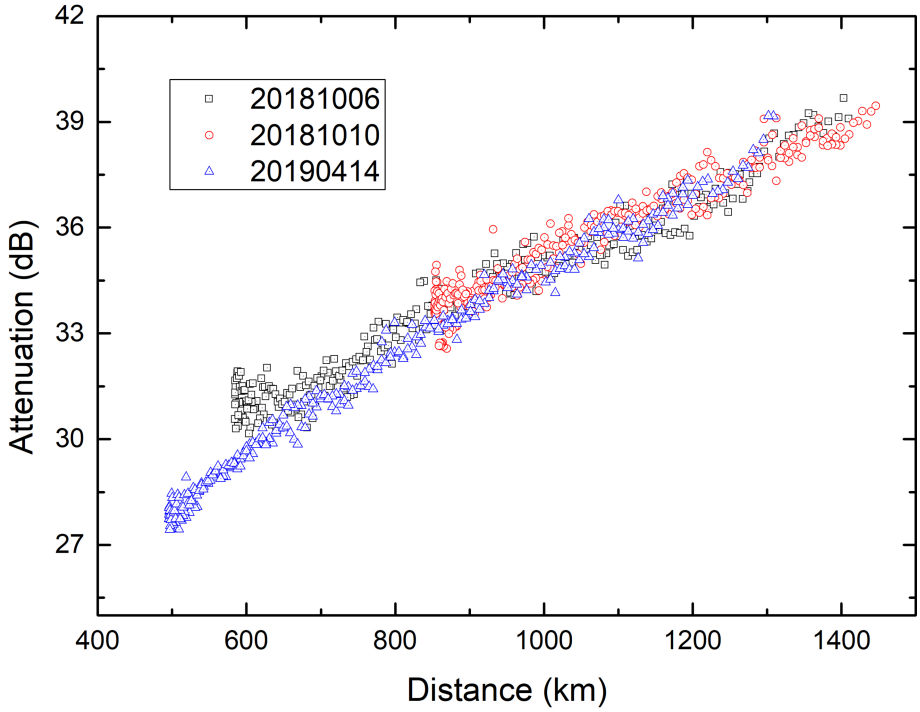
Competing interests The authors declare no competing interests.

Additional information

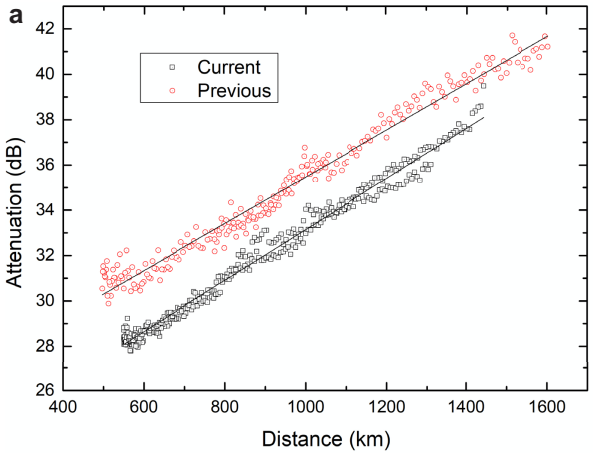
Correspondence and requests for materials should be addressed to J.-W.P. or C.-Z.P.



Extended Data Fig. 1 | Satellite-to-Delingha link efficiencies under different weather conditions. a, The data in previous work[23] was taken in different orbits during the period of 7 December 2016 to 22 December 2016. **b**, The data in current work was taken in different orbits during the period of 6 September 2018 to 22 October 2018. Here the change of link efficiencies on different days was caused by the weather conditions.



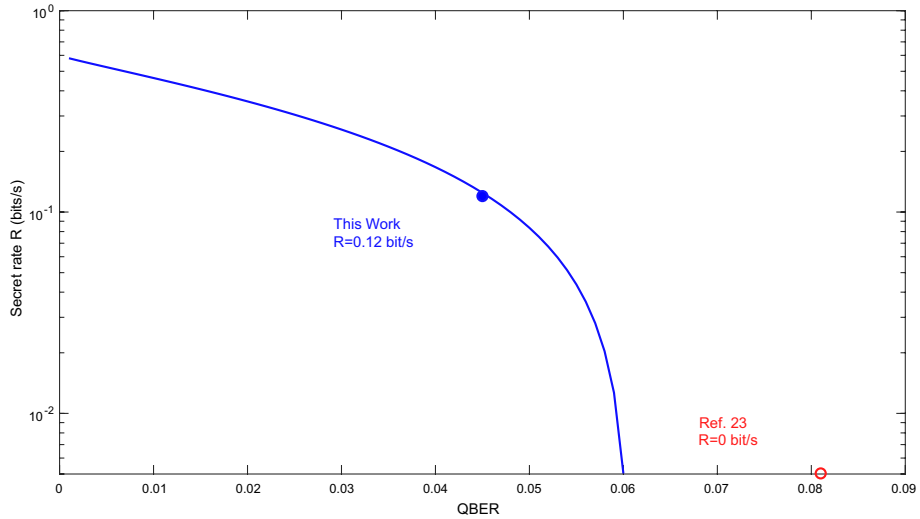
Extended Data Fig. 2 | Multiple orbits of satellite-to-Delingha link efficiencies under good weather conditions. Stable and high collection efficiencies were observed during the period of October 2018 to April 2019.



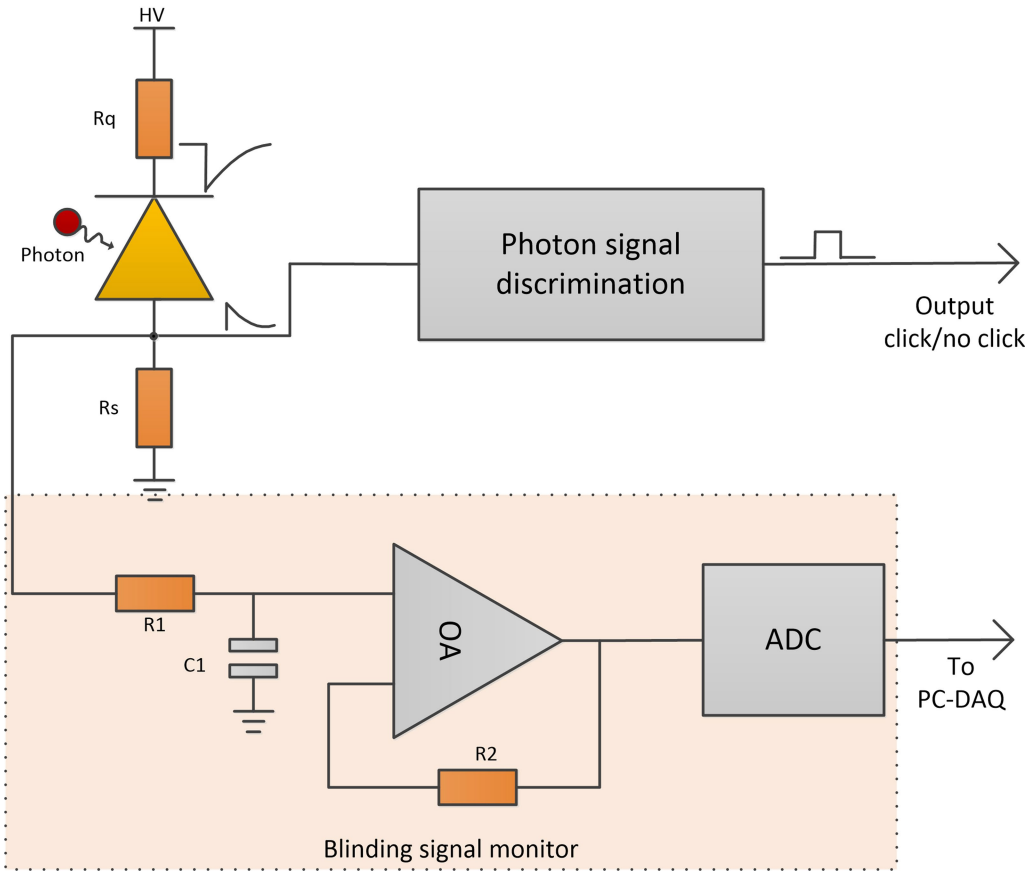
b

Distance (km)	Previous (dB)	Current (dB)	Improvement (dB)
600	31.63 ± 0.04	28.63 ± 0.03	3.00 ± 0.05
800	33.56 ± 0.06	30.94 ± 0.04	2.62 ± 0.07
1000	35.96 ± 0.07	32.97 ± 0.05	2.99 ± 0.09
1200	37.21 ± 0.08	34.84 ± 0.06	2.37 ± 0.1

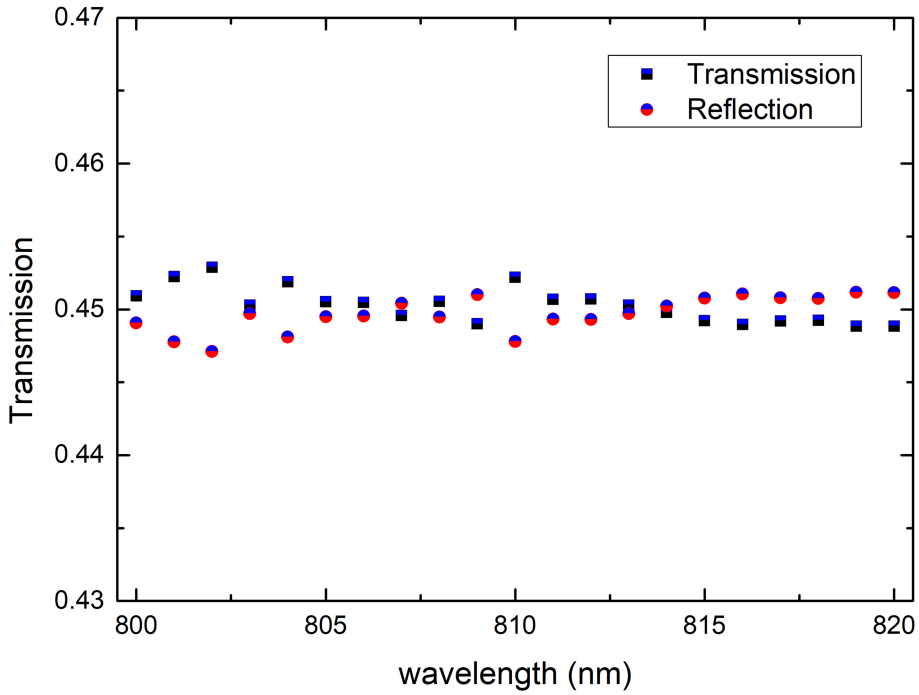
Extended Data Fig. 3 | The comparison of satellite-to-Delingha link efficiency under the best-orbit condition. a, After improving the link efficiency with high-efficiency telescopes and follow-up optics, on average, the current work shows a 3-dB enhancement in the collection efficiency over that of ref. [23]. The lines are linear fits to the data. **b**, Some representative values.



Extended Data Fig. 4 | The finite-key secret key rate *R* versus the QBER. For the 3,100 s of data collected in our experiment, a QBER of below about 6.0% is required to produce a positive key. The previous work[23] demonstrated a QBER of 8.1%, which is not sufficient to generate a secret key. In this work, a QBER of 4.5% and a secret key rate of 0.12 bits per second are demonstrated over 1,120 km. If one ignores the important finite-key effect, the QBER in ref. [23] is slightly lower than the well known asymptotic limit of 11% (ref. [43]).



Extended Data Fig. 5 | Schematics of the detection and blinding-attack monitoring circuit. The biased voltage (HV) is applied to an avalanche photodiode through a passive quenching resistance (R_q = 500 kΩ) and a sampling resistance (R_s = 10 kΩ). The avalanche signals are read out as click or no-click events through a signal-discrimination circuit. The blinding signal monitor is shown in the dot-dash diagram. A resistor-capacitor filter and a voltage follower are used to smooth and minimize the impact on the signals. The outputs of an analogue to digital converter (ADC), at a sampling rate of 250 kHz, are registered by computer data acquisition (PC-DAQ). R1, resistor; C1, capacitor; OA, operational amplifier.



Extended Data Fig. 6 | The transmission of the beam splitter within the selected bandwidth of wavelength.

Extended Data Table 1 | Parameters of the system detection efficiencies

Serial number	Current (A)	Dark counts (cps)	Efficiency (%)
B5213	0.956	55	53.31
B4973	0.9	207	53.64
B4976	1.064	60	53.16
B5214	0.984	55	53.16
B4972	0.966	32	53.78
B4974	0.929	55	53
B4977	1.067	64	53.16
B4978	0.965	26	53.16

cps, counts per second.

Extended Data Table 2 | Comparison of the results between this work and the earlier experiment[23]

	Time (orbit)	Coincidence counts	Fidelity	QBER	S
Earlier experiment [23]	250s (1)	268	0.869 ± 0.085	6.55% ± 4.25% (estimate)	
	1059s (6)	1167	--	8.10% ± 1.59% (estimate)	2.37 ± 0.09
This work	226s (1)	1021	--	4.74% ± 1.23% (estimate)	2.56 ± 0.07
	3100s	6200	0.910 ± 0.007 (estimate)	4.51% ± 0.37%	--

S, Bell parameter.

Extended Data Table 3 | Typical quantum attacks and our countermeasures

Attack	Brief Description	Countermeasure
Detector efficiency mismatch[27, 37, 38]	Eve exploits efficiency mismatch to control detectors	Free-running detectors
Detector blinding[26, 39]	Eve manipulates the detectors by sending bright light	Monitoring electronics
Detector damage[40]	Eve sends ultra-strong light to damage the detectors	Monitoring electronics
Detector dead-time[28]	Eve controls the detector by exploiting dead time	Free-running detectors
Beam-splitter[29]	Beam-splitter ratio is wavelength-dependent	Frequency filter
Spatial-mode[30]	Detectors have efficiency mismatch in spatial domain	Spatial filter

Data are from refs. [26–30,37–40].

Extended Data Table 4 | Measured correlation coefficients required for the CHSH inequality

E (φ_1, φ_2)	(0°, 22.5°)	(0°, 67.5°)	(45°, 22.5°)	(45°, 67.5°)
Value	-0.700	0.612	-0.700	-0.544
Deviation	0.0415	0.046	0.047	0.060

E, joint polarization correlation; φ_1 and φ_2, measurement angles of Delingha and Nanshan ground stations, respectively.

Optical Fourier surfaces

https://doi.org/10.1038/s41586-020-2390-x

Received: 18 December 2019

Accepted: 31 March 2020

Published online: 24 June 2020

Check for updates

Nolan Lassaline[1], Raphael Brechbühler[1], Sander J. W. Vonk[1,2], Korneel Ridderbeek[1], Martin Spieser[3], Samuel Bisig[3], Boris le Feber[1], Freddy T. Rabouw[1,2] & David J. Norris[1]✉

Gratings[1] and holograms[2] use patterned surfaces to tailor optical signals by diffraction. Despite their long history, variants with remarkable functionalities continue to be developed[3,4]. Further advances could exploit Fourier optics[5], which specifies the surface pattern that generates a desired diffracted output through its Fourier transform. To shape the optical wavefront, the ideal surface profile should contain a precise sum of sinusoidal waves, each with a well defined amplitude, spatial frequency and phase. However, because fabrication techniques typically yield profiles with at most a few depth levels, complex 'wavy' surfaces cannot be obtained, limiting the straightforward mathematical design and implementation of sophisticated diffractive optics. Here we present a simple yet powerful approach to eliminate this design–fabrication mismatch by demonstrating optical surfaces that contain an arbitrary number of specified sinusoids. We combine thermal scanning-probe lithography[6–8] and templating[9] to create periodic and aperiodic surface patterns with continuous depth control and sub-wavelength spatial resolution. Multicomponent linear gratings allow precise manipulation of electromagnetic signals through Fourier-spectrum engineering[10]. Consequently, we overcome a previous limitation in photonics by creating an ultrathin grating that simultaneously couples red, green and blue light at the same angle of incidence. More broadly, we analytically design and accurately replicate intricate two-dimensional moiré patterns[11,12], quasicrystals[13,14] and holograms[15,16], demonstrating a variety of previously unattainable diffractive surfaces. This approach may find application in optical devices (biosensors[17], lasers[18,19], metasurfaces[4] and modulators[20]) and emerging areas in photonics (topological structures[21], transformation optics[22] and valleytronics[23]).

A patterned optical surface can be described as a Fourier sum of sinusoidal waves. Each component represents a specific spatial frequency ($g = 2\pi/\Lambda$ with period Λ) that interacts with an impinging beam. For applications, diffractive surfaces should ideally contain only the frequencies of interest. However, they are typically obtained by etching patterns into surfaces to a fixed depth, creating arrays of vertical elements (trenches, holes and pillars) dictated by fabrication rather than design. This not only contributes unwanted spatial frequencies, complicating the optical response, but restricts the number of desired Fourier components that can be included. Appropriate placement of the elements (for example, aperiodically[10,13,14,18]) can offer some additional control. Alternatively, the collective response from arrays of smaller elements—nanoscale, subwavelength resonators—can be exploited in metasurfaces[24]. However, no approach has yet offered complete control over the Fourier components in a diffractive surface. If such an approach were available, simple analytical formulas could immediately specify the sum of sinusoids needed to obtain a complex desired output.

Wavy surfaces are in principle achievable using greyscale lithography[25], which spatially adjusts the exposure of a polymeric resist to produce patterns with multiple depth levels. The surface profile can then be transferred into the underlying substrate via etching. However, greyscale lithography has not yet provided sufficient spatial resolution or depth control to create arbitrary optical surfaces. Similarly, interference lithography, which exposes the resist to multiple overlapping optical beams, can generate complex diffractive surfaces[26,27]. But they contain at most a few spatial frequencies, constrained by the exposure wavelengths.

To obtain arbitrary control over the Fourier components, we first designed our structure by taking the Fourier transform of the desired diffraction pattern. After converting this analytical function into a two-dimensional (2D) greyscale bitmap (8-bit depth with 10 nm × 10 nm pixels; see Methods and Extended Data Fig. 1), we then use thermal scanning-probe lithography[6–8] to raster-scan a heated cantilever with a sharp tip across a polymer film, locally removing material to match the bitmap depth at each pixel. The simultaneous monitoring of the surface topography by the tip for feedback means that arbitrary surfaces with sub-nanometre depth control and high spatial resolution (<100 nm) can be written at about 6 s μm^{-2}. These profiles can provide diffractive elements directly or be used as an etch mask or template. We exploit templating to replicate the pattern in other materials[9].

Figure 1 demonstrates our approach with sinusoidal gratings modulated in one dimension (1D, periodic in x, constant in y), templated into silver (Ag), with one, two or three Fourier components

[1]Optical Materials Engineering Laboratory, Department of Mechanical and Process Engineering, ETH Zurich, Zurich, Switzerland. [2]Debye Institute for Nanomaterials Science, Utrecht University, Utrecht, The Netherlands. [3]Heidelberg Instruments Nano/SwissLitho, Zurich, Switzerland. ✉e-mail: dnorris@ethz.ch

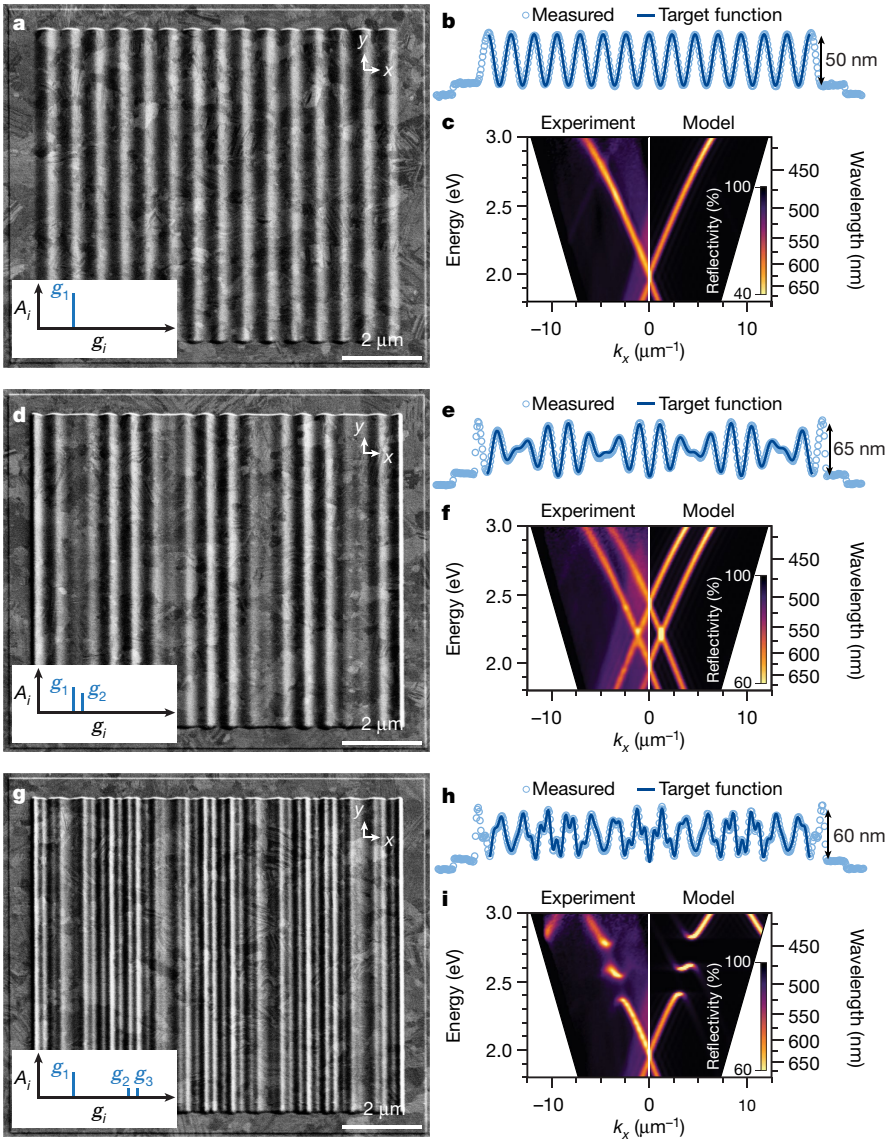


Fig. 1 | Fourier surfaces modulated in one dimension. a, **d**, **g**, Scanning-electron micrographs (SEMs, 30° tilt) of Ag gratings with 1, 2 or 3 sinusoidal components. The insets show the sinusoidal amplitudes A_i and spatial frequencies g_i. All design parameters are given in Extended Data Table 1. **b**, **e**, **h**, Measured (atomic force microscopy) and targeted surface topographies for the structures in **a**, **d** and **g**. Scan lengths are 11.3 μm and represent a single line in the structures. All target functions account for a slight distance miscalibration in the thermal scanning probe. The measured RMS error for the patterns are 1.8 nm, 2.1 nm and 2.3 nm, respectively (see Methods). **c**, Experimental (left) and modelled (right) angle-resolved reflectivity spectra (energy versus in-plane photon wavevector along the grating, k_x, with $k_y \approx 0$) for the structure in **a**. The orange lines represent decreased reflectivity at photon angles that launch surface plasmon polaritons (SPPs). These lines trace the SPP dispersion, displaced into the light cone by g_1. The black region represents energies and angles accessible in experiment (Extended Data Fig. 3). **f**, The two-component grating provides two photon–SPP coupling channels, doubling the orange lines. **i**, The three-component grating was designed to exhibit two plasmonic stopbands.

(Fig. 1a, d, g). Ag was chosen as a low-loss metal to explore non-transparent reflective gratings. The insets show the targeted amplitudes, A_i, and spatial frequencies, g_i, for sinusoid i (see the Methods for the analytical formulas for all surfaces). Because our structures are finite in size, their Fourier spectra will be slightly broader than in the analytical design (see modelling in Methods). The measured topographies for the patterns (Fig. 1b, e, h) show that the process faithfully reproduces the targeted profile with 1.8–2.3 nm root-mean-square (RMS) error (see Methods and Extended Data Fig. 2). These low values indicate that the desired Fourier components are dominant. Indeed, a detailed analysis for the single sinusoid (Extended Data Fig. 2) shows that the second harmonic is the largest error component with an amplitude of only 3.5% of A_1 (0.9 nm).

To test the optical response of our gratings, we measure angle-resolved reflectivity spectra by imaging the back focal plane of an optical microscope onto a spectrometer[28,29] (Methods; Extended Data Fig. 3a). Each sinusoidal component (here periodic in x) can provide momentum $\mathbf{g}_i = (2\pi/\Lambda_i)\hat{\mathbf{x}}$ (where $\hat{\mathbf{x}}$ is the unit vector along x) to an impinging beam. These contributions can affect the outgoing angle of the radiation or lead to electromagnetic surface waves—surface plasmon polaritons (SPPs)—that propagate along the Ag–air interface with in-plane wavevector $\mathbf{k}_{SPP}$. We use the latter process (photon–SPP coupling) to characterize the capabilities of our surfaces.

We measure reflectivity as a function of the in-plane wavevector $\mathbf{k}_{\parallel}$ of the incoming light. Figure 1c plots results for the single-sinusoidal grating for $\mathbf{k}_{\parallel} = k_x\hat{\mathbf{x}}$ (that is, energy versus k_x with $k_y \approx 0$; see Extended Data Fig. 3b). A linear polarizer was used to select only p-polarized light, which couples to SPPs (Methods). Decreased reflectivity (orange lines) occurs when $\mathbf{k}_{\parallel} \pm \mathbf{g}_1 = \mathbf{k}_{SPP}$. Thus, the grating creates a photon–SPP coupling channel, allowing the plasmonic dispersion to be optically

probed. The match between the data and our analytical model (Fig. 1c; Methods), both here and below, confirms the fidelity of our process. See also Extended Data Fig. 4.

By including additional Fourier components, increasingly complex diffractive surfaces can be constructed. With two spatial frequencies $\mathbf{g}_1$ and $\mathbf{g}_2$ (Fig. 1d, e), two photon–SPP coupling channels open (Fig. 1f). Furthermore, SPP–SPP coupling arises if one of the spatial frequencies satisfies $\mathbf{k}_{SPP} \pm \mathbf{g}_i = \mathbf{k}'_{SPP}$, where $\mathbf{k}_{SPP}$ and $\mathbf{k}'_{SPP}$ are wavevectors for SPPs propagating in different in-plane directions. This leads to a plasmonic stopband[30,31] (Extended Data Fig. 3b). Extended Data Fig. 5 shows an example at $\mathbf{k}_{\parallel} = 0$ when $\mathbf{g}_2 = 2\mathbf{g}_1$. Although we have focused so far on the spatial frequencies of the sinusoids, our fabrication approach also allows independent control of their phase and amplitude. In Extended Data Fig. 5, phase is used to render either the upper or lower stopband edge 'dark' (not coupled to photons)[30]. Extended Data Fig. 6 uses amplitude to tune the stopband width (in energy) from 0 eV to about 0.5 eV. More generally, by adding further sinusoids, more complex plasmonic dispersions can be obtained. For example, Fig. 1g shows a three-component grating that results in multiple stopbands. These can be placed at arbitrary energies and incident photon angles. Although the surface profile (Fig. 1h) would be difficult to intuit, Fourier design followed by our process leads directly to the desired response (Fig. 1i). When such Fourier surfaces are converted into conventional two-depth-level gratings, the response is corrupted by unwanted spatial frequencies (Extended Data Fig. 4).

The control of sinusoidal components, shown above for patterns modulated in 1D with all $\mathbf{g}_i$ along $\hat{\mathbf{x}}$, can be extended to patterns modulated in 2D (Extended Data Fig. 7a, b). For example, if we sum two sinusoids, one with $\mathbf{g}_1$ along $\hat{\mathbf{x}}$ and the other with $\mathbf{g}_2$ rotated by 10° from $\hat{\mathbf{x}}$, we obtain the moiré spatial interference pattern in Fig. 2a. For a 40° rotation, the pattern in Fig. 2b results. Because these gratings now provide in-plane momentum along both $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, we plot reflectivity versus in-plane wavevector components k_x and k_y, taking a fixed-energy slice from the full dispersion diagram (Extended Data Fig. 3c). The linear polarizer used in Fig. 1 was removed (Methods). The experimentally accessible wavevectors for such a 'k-space image' (due to our finite collection angle) are within the solid white circles in Fig. 2c, d. The measured reflectivity exhibits two pairs of orange arcs, each pair representing solutions to $\mathbf{k}_{\parallel} \pm \mathbf{g}_i = \mathbf{k}_{SPP}$ (Extended Data Fig. 3d). Both plots (Fig. 2c, d) also include the 2D Fourier transform of the surface profile; the Fourier components $\pm\mathbf{g}_1$ and $\pm\mathbf{g}_2$ appear as orange spots outside the white circle and quantitatively explain the measured arcs. Even for only a 10° rotation, which leads to subtle intricacies in the surface pattern (Fig. 2a), the expected diffraction is observed.

Our approach can also exploit different basis functions. Extended Data Fig. 7c, d shows a circular sinusoidal grating and a moiré interference pattern generated from two such gratings. Functions with varying local spatial frequencies can also be employed. Figure 2e shows a sinusoidal 'zone plate' (Methods). In general, such structures can act as Fresnel lenses to focus electromagnetic radiation by diffraction, representing a unit of holographic information. Here, our zone plates have dimensions appropriate for X-ray optics[32,33], with the added benefit of continuous depth control, highly desirable for this application[34].

While the number of spatial components is arbitrary, several important symmetries can be generated by combining only a few sinusoids. Figure 3a, b shows a periodic pattern created from three sinusoids with 60° rotation between them. The resulting profile is hexagonal, with sixfold rotational symmetry, a typical design for 2D arrays of holes or pillars. However, in our structure, the 2D Fourier spectrum is specified. The corresponding k-space image (Fig. 3c) reveals six orange arcs from photon–SPP coupling. Figure 3d, e shows a surface with 12-fold rotational symmetry created from six sinusoids with 30° rotation between them. In k-space, 12 orange arcs appear (Fig. 3f). This profile, which does not possess translational symmetry, would be quasiperiodic if infinitely extended. Similar photonic quasicrystals using quasiperiodic arrays

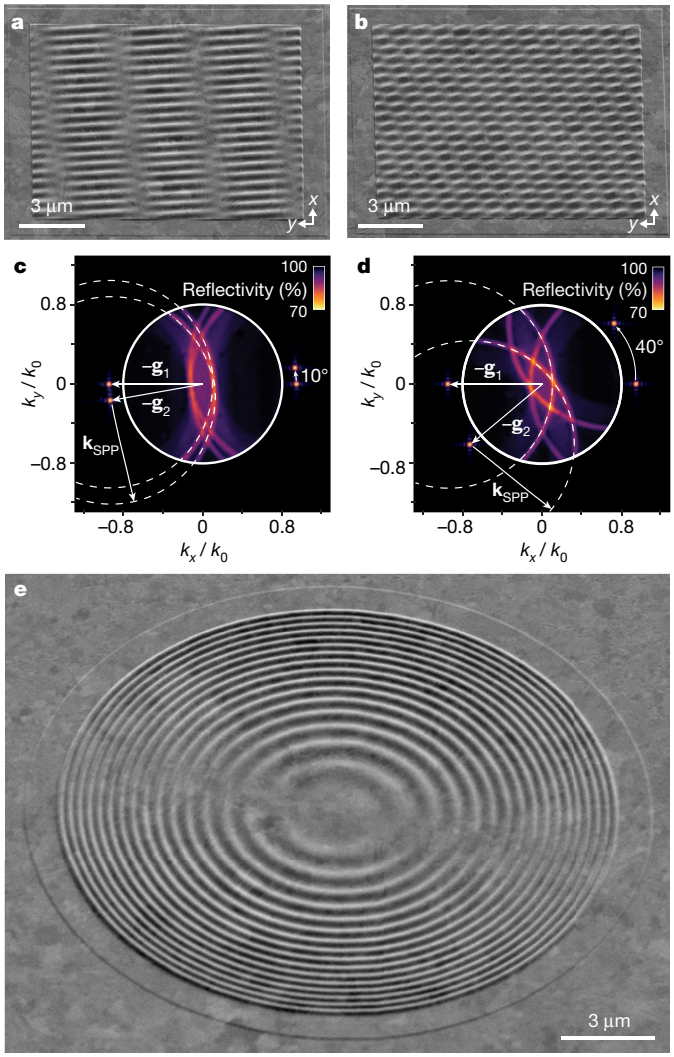


Fig. 2 | Fourier surfaces modulated in two dimensions. a, **b**, SEMs (45° tilt) of moiré patterns in Ag from two superimposed sinusoids: one with $\mathbf{g}_1$ along $\hat{\mathbf{x}}$ and the other with $\mathbf{g}_2$ rotated by 10° or 40° from $\hat{\mathbf{x}}$, respectively. See Extended Data Fig. 7. **c**, **d**, Measured k-space images (inside solid white circles) for photons (570 nm wavelength) reflected from patterns in **a** and **b**, respectively. k_x and k_y are normalized by the magnitude of the photon wavevector, k_0. Four orange arcs appear, caused by decreased reflectivity when photons launch SPPs with wavevector $\mathbf{k}_{SPP}$, that is, when $\mathbf{k}_{\parallel} \pm \mathbf{g}_i = \mathbf{k}_{SPP}$. $\pm\mathbf{g}_1$ and $\pm\mathbf{g}_2$ are shown as orange points outside the white circles. Their positions are determined from the 2D Fourier transform of the surface profiles used to define the structures. In **c** and **d**, we see that $\mathbf{k}_{\parallel} = -\mathbf{g}_2 + \mathbf{k}_{SPP}$ forms an orange arc in k-space. **e**, SEM (45° tilt) of a Ag sinusoidal zone plate. For all structural design parameters, see Extended Data Table 1.

of trenches or holes have been reported for laser applications[10,18,35]. However, optimizing their design is computationally intensive and still results in 2D Fourier spectra with many unwanted spatial frequencies. Our structures are designed with simple analytical functions and exhibit precise control over the Fourier components.

To demonstrate the utility of our approach, we address a current limitation in photonics. The push for miniaturized optical systems has led to waveguides integrated into a single thin layer that exploits diffractive optics for in- and outcoupling of light[20,36]. For these devices, multiple wavelengths should ideally be diffracted between free-space beams and propagating waveguide modes at a common angle. However, current single-spatial-frequency gratings cause them to diffract at different, highly specific angles, resulting in bigger, more complicated devices.

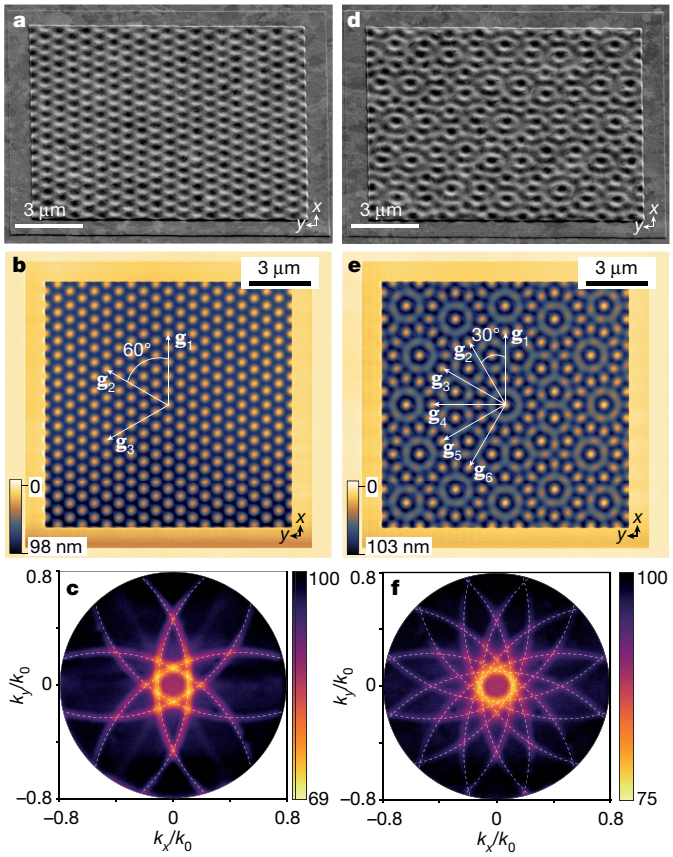


Fig. 3 | Periodic and quasiperiodic Fourier surfaces. a, **d**, SEMs (45° tilt) of periodic and quasiperiodic optical Fourier surfaces templated in Ag with 6- and 12-fold rotational symmetry, defined when three and six sinusoids are superimposed, respectively. **b**, **e**, Measured topographies (obtained during patterning) for the polymer films (PMMA/MA; see Methods) used to template the structures in **a** and **d**, respectively. All sinusoids have $\Lambda = 600$ nm and their corresponding vectors $\mathbf{g}_i$ are oriented in-plane, as shown, spaced by 60° and 30°, respectively. **c**, **f**, Measured *k*-space reflectivity images for photons (570 nm wavelength) incident on the patterns in **a** and **d**, respectively. Six and 12 orange arcs appear, caused by decreased reflectivity when photons launch SPPs with wavevector $\mathbf{k}_{SPP}$, that is, when $\mathbf{k}_{\parallel} \pm \mathbf{g}_i = \mathbf{k}_{SPP}$ (dashed white lines). k_x and k_y are normalized by the magnitude of the photon wavevector, k_0. For all structural design parameters, see Extended Data Table 1.

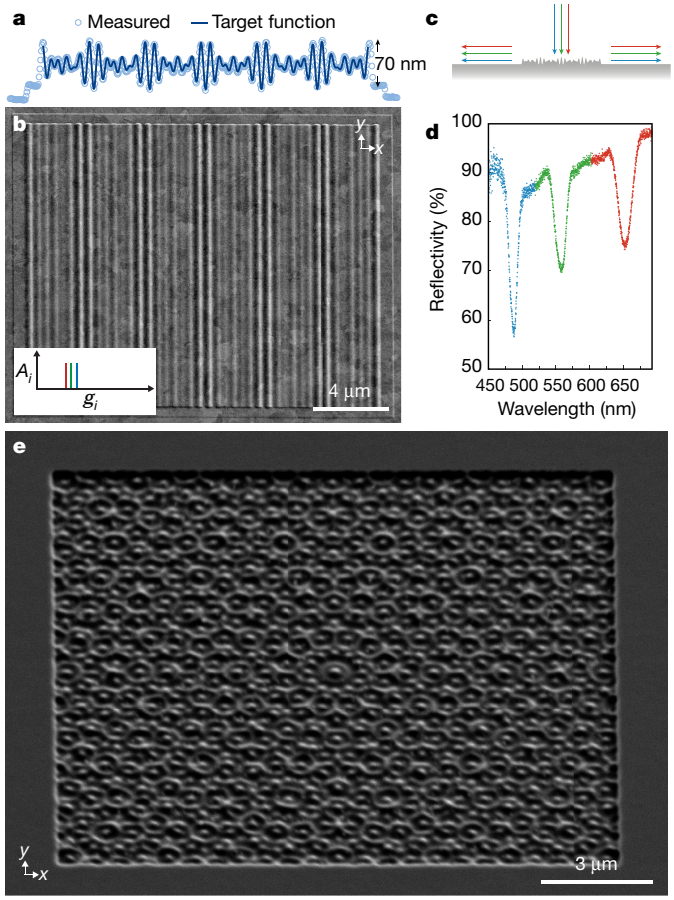


Fig. 4 | Applications of Fourier surfaces. a, Comparison of the measured (atomic force microscopy) and targeted surface topography (accounting for a slight distance miscalibration in the thermal scanning probe) for a Ag Fourier surface that couples red, green and blue photons at normal incidence to SPPs. Scan length is 22.0 μm. The profile contains three sinusoids with design periods $\Lambda_1 = 620$ nm, $\Lambda_2 = 520$ nm and $\Lambda_3 = 445$ nm. **b**, SEM (30° tilt) of the Ag Fourier surface in **a**. The inset shows the sinusoidal amplitudes A_i and spatial frequencies g_i. All design parameters are given in Extended Data Table 1. **c**, Cartoon of the coupling of red, green and blue light simultaneously at normal incidence. **d**, Measured reflectivity as a function of photon wavelength for light at normal incidence (within ±0.25°). The three prominent reflectivity dips (coloured as a visual guide) correspond to the coupling of red, green and blue light at normal incidence. **e**, SEM (45° tilt) of a 12-fold rotationally symmetric quasicrystal, defined with twelve sinusoids, etched into Si. For design parameters, see Extended Data Table 1.

With Fourier surfaces, a simple solution is immediately available. Three spatial frequencies can be included on a single surface to diffract three colours at a common angle. Figure 4a, b shows such a profile, designed, implemented and templated in Ag. The three sinusoidal components simultaneously couple red, green and blue photons at normal incidence (Fig. 4c), as seen by the three reflectivity dips in Fig. 4d, which arise due to photon–SPP coupling.

Additional applications of optical Fourier surfaces can benefit from deeper structures and a diverse material library. We patterned polymers of various refractive index (Methods) up to about 300 nm deep. When these deeper surfaces are templated into Ag, the resulting Fourier surfaces can provide efficient diffraction gratings (Extended Data Fig. 8). For a p-polarized beam at normal incidence, we measured 97 ± 5% in the +1 and −1 diffraction orders for a single-component sinusoidal grating. With two sinusoidal components, a 'blazed' Fourier surface is obtained that diffracts nearly all intensity into just the +1 diffraction order. The polymer profiles can also be transferred into substrates via etching, for example silicon (Si; Fig. 4e) or silicon nitride (SiN_x; Extended Data Fig. 9). With this we could amplify the profile depth[8]. Like the patterned polymer, the etched substrate can provide a multi-use template[9]. Extended Data Fig. 9 shows a titania (TiO_2) Fourier surface templated from an etched Si substrate.

Thus, 'wavy' diffractive surfaces can be provided for a broad spectral range (X-ray to infrared). Templating, extendable to rollable substrates[37], enables high-throughput production of many materials including active and multilayer solids[38,39]. Optical wavefronts can be manipulated (including direction, phase and polarization[40,41]) with diffractive surfaces that can be accurately placed within or on top of elements in integrated photonic devices, allowing miniaturized optical systems[20,36]. Thus, researchers in photonics can exploit the previously unavailable capabilities of optical Fourier surfaces to address applications as well as to explore emerging phenomena.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2390-x.

1. Hopkinson, F. & Rittenhouse, D. An optical problem, proposed by Mr. Hopkinson, and solved by Mr. Rittenhouse. *Trans. Am. Phil. Soc.* **2**, 201–206 (1786).
2. Gabor, D. A new microscopic principle. *Nature* **161**, 777–778 (1948).
3. Ebbesen, T. W., Lezec, H. J., Ghaemi, H. F., Thio, T. & Wolff, P. A. Extraordinary optical transmission through sub-wavelength hole arrays. *Nature* **391**, 667–669 (1998).
4. Khorasaninejad, M. et al. Metalenses at visible wavelengths: diffraction-limited focusing and subwavelength resolution imaging. *Science* **352**, 1190–1194 (2016).
5. Goodman, J. W. *Introduction to Fourier Optics* (W. H. Freeman, 2017).
6. Mamin, H. J. & Rugar, D. Thermomechanical writing with an atomic force microscope tip. *Appl. Phys. Lett.* **61**, 1003–1005 (1992).
7. Pires, D. et al. Nanoscale three-dimensional patterning of molecular resists by scanning probes. *Science* **328**, 732–735 (2010).
8. Rawlings, C. D. et al. Control of the interaction strength of photonic molecules by nanometer precise 3D fabrication. *Sci. Rep.* **7**, 16502 (2017).
9. Nagpal, P., Lindquist, N. C., Oh, S. H. & Norris, D. J. Ultrasmooth patterned metals for plasmonics and metamaterials. *Science* **325**, 594–597 (2009).
10. Blanchard, R. et al. Gratings with an aperiodic basis: single-mode emission in multi-wavelength lasers. *New J. Phys.* **13**, 113023 (2011).
11. Sunku, S. S. et al. Photonic crystals for nano-light in moiré graphene superlattices. *Science* **362**, 1153–1156 (2018).
12. Wang, P. et al. Localization and delocalization of light in photonic moiré lattices. *Nature* **577**, 42–46 (2020).
13. Matsui, T., Agrawal, A., Nahata, A. & Vardeny, Z. V. Transmission resonances through aperiodic arrays of subwavelength apertures. *Nature* **446**, 517–521 (2007).
14. Martins, E. R. et al. Deterministic quasi-random nanostructures for photon control. *Nat. Commun.* **4**, 2665 (2013).
15. Ozaki, M., Kato, J.-i. & Kawata, S. Surface-plasmon holography with white-light illumination. *Science* **332**, 218–220 (2011).
16. Zheng, G. et al. Metasurface holograms reaching 80% efficiency. *Nat. Nanotechnol.* **10**, 308–312 (2015).
17. Tittl, A. et al. Imaging-based molecular barcoding with pixelated dielectric metasurfaces. *Science* **360**, 1105–1109 (2018).
18. Mahler, L. et al. Quasi-periodic distributed feedback laser. *Nat. Photon.* **4**, 165–169 (2010).
19. Yoshida, M. et al. Double-lattice photonic-crystal resonators enabling high-brightness semiconductor lasers with symmetric narrow-divergence beams. *Nat. Mater.* **18**, 121–128 (2019).
20. Ayata, M. et al. High-speed plasmonic modulator in a single metal layer. *Science* **358**, 630–632 (2017).
21. Lu, L., Joannopoulos, J. D. & Soljacic, M. Topological photonics. *Nat. Photon.* **8**, 821–829 (2014).
22. Pendry, J. B., Huidobro, P. A., Luo, Y. & Galiffi, E. Compacted dimensions and singular plasmonic surfaces. *Science* **358**, 915–917 (2017).
23. Hu, G. et al. Coherent steering of nonlinear chiral valley photons with a synthetic Au–WS_2 metasurface. *Nat. Photon.* **13**, 467–472 (2019).
24. Genevet, P., Capasso, F., Aieta, F., Khorasaninejad, M. & Devlin, R. Recent advances in planar optics: from plasmonic to dielectric metasurfaces. *Optica* **4**, 139–152 (2017).
25. Kim, J., Joy, D. C. & Lee, S. Y. Controlling resist thickness and etch depth for fabrication of 3D structures in electron-beam grayscale lithography. *Microelectron. Eng.* **84**, 2859–2864 (2007).
26. Dakss, M. L., Kuhn, L., Heidrich, P. F. & Scott, B. A. Grating coupler for efficient excitation of optical guided waves in thin films. *Appl. Phys. Lett.* **16**, 523–525 (1970).
27. Campbell, M., Sharp, D. N., Harrison, M. T., Denning, R. G. & Turberfield, A. J. Fabrication of photonic crystals for the visible spectrum by holographic lithography. *Nature* **404**, 53–56 (2000).
28. Shi, L. et al. Spatial coherence properties of organic molecules coupled to plasmonic surface lattice resonances in the weak and strong coupling regimes. *Phys. Rev. Lett.* **112**, 153002 (2014).
29. Kurvits, J. A., Jiang, M. & Zia, R. Comparative analysis of imaging configurations and objectives for Fourier microscopy. *J. Opt. Soc. Am. A* **32**, 2082–2092 (2015).
30. Barnes, W. L., Preist, T. W., Kitson, S. C. & Sambles, J. R. Physical origin of photonic energy gaps in the propagation of surface plasmons on gratings. *Phys. Rev. B* **54**, 6227–6244 (1996).
31. Joannopoulos, J. D., Johnson, S. G., Winn, J. N. & Meade, R. D. *Photonic Crystals: Molding the Flow of Light* (Princeton Univ. Press, 2008).
32. Chao, W., Harteneck, B. D., Liddle, J. A., Anderson, E. H. & Attwood, D. T. Soft X-ray microscopy at a spatial resolution better than 15 nm. *Nature* **435**, 1210–1213 (2005).
33. Wang, Y., Yun, W. & Jacobsen, C. Achromatic Fresnel optics for wideband extreme-ultraviolet and X-ray imaging. *Nature* **424**, 50–53 (2003).
34. Di Fabrizio, E. et al. High-efficiency multilevel zone plates for keV X-rays. *Nature* **401**, 895–898 (1999).
35. Vitiello, M. S. et al. Photonic quasi-crystal terahertz lasers. *Nat. Commun.* **5**, 5884 (2014).
36. Huang, Z. Q., Marks, D. L. & Smith, D. R. Out-of-plane computer-generated multicolor waveguide holography. *Optica* **6**, 119–124 (2019).
37. Yoo, D., Johnson, T. W., Cherukulappurath, S., Norris, D. J. & Oh, S. H. Template-stripped tunable plasmonic devices on stretchable and rollable substrates. *ACS Nano* **9**, 10647–10654 (2015).
38. Wuttig, M., Bhaskaran, H. & Taubner, T. Phase-change materials for non-volatile photonic applications. *Nat. Photon.* **11**, 465–476 (2017).
39. Shaltout, A. M., Shalaev, V. M. & Brongersma, M. L. Spatiotemporal light control with active metasurfaces. *Science* **364**, eaat3100 (2019).
40. Arbabi, A., Horie, Y., Bagheri, M. & Faraon, A. Dielectric metasurfaces for complete control of phase and polarization with subwavelength spatial resolution and high transmission. *Nat. Nanotechnol.* **10**, 937–943 (2015).
41. Rubin, N. A., D'Aversa, G., Chevalier, P., Shi, Z., Chen, W. T. & Capasso, F. Matrix Fourier optics enables a compact full-Stokes polarization camera. *Science* **365**, eaax1839 (2019).

Methods

Fourier-surface design

All surfaces were designed using analytical functions. In general, 1D real-space height profiles, $f(x)$, can be obtained from the desired Fourier spectrum, $F(K)$, via the 1D inverse Fourier transform:

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(K)\mathrm{e}^{\mathrm{i}Kx}\mathrm{d}K \quad (1)$$

K is a spatial-frequency variable and $F(K)$ describes the spatial frequencies (g) contained in the surface profile. Similarly, 2D height profiles, $f(x, y)$, follow from the 2D inverse Fourier transform of $F(K_x, K_y)$:

$$f(x,y) = \frac{1}{(2\pi)^2}\iint_{-\infty}^{\infty} F(K_x, K_y)\mathrm{e}^{\mathrm{i}(K_x x + K_y y)}\mathrm{d}K_x\mathrm{d}K_y \quad (2)$$

K_x and K_y are spatial-frequency variables along the x and y axes. For $f(x)$ and $f(x, y)$, the origin is placed in the middle of the pattern for both x and y. All functions are defined for the pattern in the polymer surface, where x and y lie in-plane and z is perpendicular. In these formulas, the height of the surface is defined relative to the unpatterned flat surface where $z = 0$. Note that the Fourier spectra in equations (1) and (2), used to calculate the infinitely extended real-space surface profiles, neglect finite-size effects. The finite dimensions of the experimental profile lead to broadening of the Fourier spectra (see Methods section 'Analytical model').

For the Fourier surfaces in Figs. 1 and 4 and Extended Data Figs. 1, 2, 4, 5, 6, 8 and 9, the Fourier spectrum is sufficiently simple (with one, two or three Fourier components, assuming infinite size in x, y) that the height profile can be written as a sum of sinusoids:

$$f(x) = \sum_i A_i \cos(g_i x + \varphi_i) - \Delta \quad (3)$$

where A_i, g_i and φ_i are the amplitude, spatial frequency, and phase, respectively, for component i. Note that in equation (3), the sinusoidal surface profiles in the polymer are vertically shifted in z by $-\Delta$. When templating is used to transfer the pattern to Ag, the surface profile is inverted and vertically shifted in z by $+\Delta$. For clarity, all parameters for our polymer surfaces are provided in Extended Data Table 1.

For the Fourier surfaces in Figs. 2a, b, 3a, d, 4e and Extended Data Figs. 7, 9, the height profile was given by:

$$f(x,y) = \sum_i A_i \cos[g_i(x\cos\theta_i + y\sin\theta_i) + \varphi_i] - \Delta \quad (4)$$

where θ_i is the in-plane rotation angle from the x axis for component i. The circular Fourier surfaces in Extended Data Fig. 7 follow:

$$f(r,\theta) = \sum_i A_i \cos(g_i|\mathbf{r} - \mathbf{r}_i| + \varphi_i) - \Delta \quad (5)$$

where r and θ are the radial distance and polar angle, respectively. $\mathbf{r}$ is the coordinate in the surface plane and is a function of r and θ. $\mathbf{r}_i$ is the centre of circular component i. The sinusoidal zone plate[42] in Fig. 2e follows the function:

$$f(r) = A\sin\left[\pi\left(\frac{r}{L}\right)^2\right] - \Delta \quad (6)$$

where A is an amplitude and L is a characteristic length.

Bitmap generation

The analytical functions defining the Fourier surfaces are converted into bitmaps. The overall dimensions in x and y are chosen for the structure, and the analytical function is mapped onto a 10 nm × 10 nm pixel grid. The normalized depth of the structure in z was assigned for each pixel by discretizing the total normalized depth to 256 levels (8-bit precision). The physical patterning depth was assigned for each pixel by inputting the maximum physical depth of the structure to the thermal scanning-probe control software (see Methods section 'Fourier-surface fabrication'), which then assigned the physical depth for each pixel based on its 8-bit depth level. The entire process flow, from analytical mathematical design to pattern transfer to an optical material, is depicted in Extended Data Fig. 1.

Materials

1-mm-thick glass microscope slides and 1-mm-thick, 2-inch-diameter and 4-inch-diameter Si(100) wafers (1–10 Ω cm resistivity) were purchased from Paul Marienfeld and Silicon Materials, respectively. Ag (1/4-inch-diameter × 1/4-inch-long pellets, 99.999%), Au (1/8-inch-diameter × 1/8-inch-long pellets, 99.999%), TiO_2 sputter targets (200 mm diameter, 99.95%), and ultraviolet-curable epoxy (OG142-95 and OG116-31) were obtained from Kurt J. Lesker, ACI Alloys, FHR Anlagenbau, and Epoxy Technology, respectively. Tungsten dimple boats ($49 \times 12 \times 0.4$ mm^3) were bought from Umicore. Two polymer resists from Allresist GmbH were used: PMMA/MA [AR-P 617, poly(methyl methacrylate-co-methacrylic acid), 33% copolymer, 3% dilution in anisol] and CSAR [AR-P 6200, containing poly(α-methylstyrene-co-methyl chloroacrylate) in anisol]. For electron-beam lithography, the CSAR resist was developed using AR 600-546 from Allresist. Silicon cantilevers for thermal scanning-probe lithography with a tip radius of ~3–5 nm were provided by SwissLitho (SL2015-2-HPL, SL2016-3-HPL, SL2018-13-HPL and SL2018-2-MBS). Hydrochloric acid (HCl, 37%), nitric acid (HNO_3, ≥65%), sulfuric acid (H_2SO_4, ≥95%), and ammonium fluoride + hydrofluoric acid etching mixture (AF 875-125) were purchased from Sigma-Aldrich. Hydrogen peroxide (H_2O_2, 30%) was obtained from VWR Chemicals. Acetone and isopropanol (IPA) were provided by the Binnig and Rohrer Nanotechnology Center (BRNC) at IBM Zurich, where the templates were fabricated.

Fourier-surface fabrication

A Si wafer was typically used as the sample substrate. It was removed from its factory packaging in the cleanroom and used directly. The polymer resist layer was spin-coated onto it using a two-step procedure. For PMMA/MA or CSAR, the resist solution was deposited on the sample surface and accelerated at 500 rpm s^{-1} to 500 rpm for 5 s. Then the PMMA/MA (CSAR) was accelerated at 2,000 rpm s^{-1} to 2,000 rpm (2,500 rpm) for a total time of 40 s. After spin-coating, the PMMA/MA (CSAR) layer was baked at 180 °C for 5 min (150 °C for 1 min). For the deeper Fourier surface structures in Extended Data Fig. 8, the PMMA/MA spin-coating and baking procedure was repeated to double the thickness of the resist layer from ~150 nm to ~300 nm.

For thermal scanning-probe lithography, the substrate/polymer stack was placed in a NanoFrazor Explore (SwissLitho). A cantilever with a sharp tip was loaded into the cantilever holder, which was then attached to the NanoFrazor scan head. The tip was brought close to the sample and an auto-approach function achieved surface contact. The tip position, temperature response and sample tilt were calibrated. The temperature at the base of the tip was set to an initial value between 700 °C and 950 °C, depending on the cantilever model. Calibration scans were performed to optimize the patterning depth of the sinusoidal structures. The bitmap defining the desired Fourier surface was then loaded into the NanoFrazor software. The tip was scanned across the patterning surface on a 10 nm × 10 nm pixel grid. A force pulse (~6 μs) was applied at each pixel to match the depth level of the bitmap in the polymer resist. As the tip patterned the surface, it simultaneously measured the topography as in contact-mode atomic force microscopy (AFM). The measured error between the written pattern and the desired pattern was passed to a feedback loop such that the write

force could be adjusted to reach the desired depth level, if necessary. The scan progressed until all pixels in the design were patterned into the surface, at which point the tip was available to write the next pattern.

To obtain Ag diffractive surfaces, Ag was thermally evaporated[43] (Kurt J. Lesker, Nano36) onto the patterned polymer film at a pressure of about 3×10^{-7} mbar. A tungsten boat loaded with Ag pellets was heated to deposit at a rate of 25 Å s^{-1}. The process was stopped when the film thickness was around 750 nm. A glass slide was then affixed with ultraviolet-curable epoxy (OG142-95) onto the exposed Ag surface, and the glass/epoxy/Ag stack peeled off, revealing a Ag surface with the negative of the initial pattern in the polymer surface.

SiN_x surfaces were obtained by using a $Si/SiO_2/SiN_x$ stack as a substrate. A layer of SiO_2 2,000 nm thick was thermally grown onto a Si wafer, followed by chemical vapour deposition of a layer of SiN_x 200 nm thick. The wafer was diced into 1.5 cm × 1.5 cm pieces for thermal scanning-probe lithography using PMMA/MA as the polymer. The pattern in the polymer film was transferred into the underlying SiN_x substrate via reactive-ion etching (Oxford Instruments, NPG 80) using a gas mixture of 50 standard cubic centimetres per minute (sccm) CHF_3 and 5 sccm O_2. The etching was performed at a chamber pressure of 55 mTorr, with 100 W radio-frequency power and a SiN_x etch rate of 45 nm min^{-1} for 5 min, where the depth of the transferred pattern in SiN_x was approximately the same as the depth in the polymer pattern (approximately 1:1 selectivity). Afterwards, the substrate was ultrasonicated in acetone, followed by isopropanol, and blown dry with N_2.

To obtain Si surfaces for either direct use or for templating, the pattern in the polymer film was transferred into the underlying Si substrate via inductively coupled plasma etching (Oxford Instruments, Plasma Pro) using a gas mixture of 17.0 sccm SF_6, 17.5 sccm C_4F_8 and 60 sccm Ar. The Si etching was done at a chamber pressure of 20 mTorr, with a forward power of 50 W, and at a rate of ~25 nm min^{-1} for 6.33 min, where the depth of the transferred pattern in Si was approximately the same as the depth in the polymer pattern (approximately 1:1 selectivity). After etching, the sample was sonicated for 2 min in acetone and 2 min in IPA, followed by 5 min of O_2 plasma cleaning at 600 W.

Patterned TiO_2 samples were obtained by using patterned Si templates. A 25-nm-thick Au layer was thermally evaporated onto the patterned Si wafer at a pressure of approximately 3×10^{-7} mbar and a rate of 10 Å s^{-1}. TiO_2 was then radio-frequency-sputtered onto the exposed gold surface (von Ardenne, CS 320 S) with 400 W, a chamber pressure of 4×10^{-3} mbar, and 14 sccm Ar, for 160 min, resulting in an approximately 300-nm-thick film. A glass slide was then affixed with ultraviolet-curable epoxy (OG116-31) onto the exposed TiO_2 layer, and the glass/epoxy/TiO_2/Au stack peeled off, revealing a TiO_2/Au surface with the negative of the initial pattern in the Si surface. Finally, the Au layer was removed by immersing the sample in aqua regia (4:1 mixture of $HCl:HNO_3$) for 5 min. Afterwards, the sample was rinsed in deionized water and blown dry with N_2.

Binarized-surface design and fabrication

For each Fourier surface in Fig. 1 with height profile $f(x)$ (see Extended Data Table 1), a binarized version was fabricated by electron-beam lithography and etching, followed by templating (see Extended Data Fig. 4). The binarization followed a published thresholding procedure[44]. This required the electron-beam lithography resist to be exposed wherever $f(x) + \Delta < 0$. The Si substrate was then etched in these locations.

To prepare the samples, 2 × 2 cm chips (diced from a 4-inch-diameter, 1-mm-thick Si wafer) were cleaned by sonicating for 2 min in acetone and 2 min in IPA, followed by 5 min of O_2 plasma cleaning at 600 W. CSAR (electron-beam lithography resist) was deposited on the sample surfaces and accelerated at 500 rpm s^{-1} to 500 rpm for 5 s in a first spin-coating step. In a second step, the samples were accelerated at 2,000 rpm s^{-1} to 2,000 rpm for a total time of 40 s. After spin-coating, the samples were baked at 150 °C for 1 min. The samples were then loaded into an electron-beam lithography system (Vistec, NFL 5) and patterned by exposing the resist layer where specified by the thresholding procedure[44]. After exposure, the samples were developed in AR 600-546 for 1 min, and subsequently rinsed in IPA. The patterns were etched to depths approximately matching that of the corresponding Fourier surfaces from Fig. 1 with HBr-based reactive-ion etching (Oxford, Plasmalab System100). The Si etching was done using 40 sccm HBr at a chamber pressure of 3 mTorr, with a forward power of 200 W, radio-frequency power of 20 W, and at a rate of approximately 30 nm min^{-1} for 2 min. After etching, the resist was removed by subsequent sonication in acetone and in IPA, followed by 2 min of O_2 plasma cleaning at 600 W, a dip in buffered hydrofluoric acid (1:7 mixture of AF 875-125 and H_2O), and rinsing with H_2O. The samples were cleaned in piranha (1:1 mixture of H_2SO_4:H_2O_2) for 15 min, ultrasonicated in H_2O and in IPA, and blown dry with N_2. (Caution: care should be taken with piranha as it reacts violently with solvents and other organic materials.) These binarized surfaces were then replicated in Ag using the same procedure as for the Fourier surfaces.

Surface-topography characterization

The topography of the Fourier surfaces was measured by the scanning probe during patterning and independently verified with AFM on the templated Ag surface. The topography of our Ag single-sinusoidal surface (Fig. 1a, b) is analysed in Extended Data Fig. 2. AFM scans (Bruker, Dimension FastScan AFM with a Bruker NCHV-A cantilever) were collected in tapping mode under ambient conditions. The raw data was processed by first removing the instrumental high-frequency scan noise in the scanning-probe analysis software Gwyddion (version 2.54, http://gwyddion.net). Next, row alignment and plane-levelling were performed in MATLAB (version 2019a, http://ch.mathworks.com/products/matlab.html) to obtain the corrected data, shown in Extended Data Fig. 2a. These data were then analysed by fitting a sinusoidal function (with the form shown in Extended Data Table 1 for Fig. 1a; periodic along x, constant in y), where the fit parameters and residuals were extracted. The amplitude and period of the fitted function were A_1 = 25.5 nm (2% larger than design value) and Λ = 610 nm (1.7% larger than design value), respectively. As we obtained a consistent horizontal distance error in both our etched and templated gratings, we attributed this error to a distance miscalibration in the thermal scanning probe. The RMS error between the design function and measured topography for the structure in Fig. 1a was found to be 1.8 nm after this error was taken into account. A similar procedure was used to extract RMS errors for other Fourier surfaces, as reported in the legends of Fig. 1 and Extended Data Fig. 9. See Extended Data Fig. 2 for further details. For the photonic diffraction gratings in Extended Data Fig. 8, a slight nonlinearity in the patterning of deeper structures was also taken into account.

Optical characterization

The optical-characterization setup is depicted in Extended Data Fig. 3a. Ag surfaces were measured with an inverted optical microscope (Nikon, Eclipse Ti-U) equipped with a 50× air objective (Nikon, TU Plan Fluor, numerical aperture NA = 0.8). A halogen lamp was used to illuminate the sample. The lamp filament was imaged onto the back focal plane of the microscope objective. After a beamsplitter, the light was focused onto the sample and then collected by the same objective. Reflected light was transmitted through the beamsplitter and passed through a circular aperture in the real-space image plane to isolate the structure of interest. The back focal plane was imaged onto the entrance slit of an imaging spectrograph (Andor Shamrock 303i) where it was relayed to a sensitive digital camera (Andor Zyla PLUS sCMOS) for image acquisition. Reflectivity measurements were obtained for both dispersed k-space measurements (Fig. 1c, f, i, Extended Data Fig. 4b, d, f, Extended Data Fig. 5b and Extended Data Fig. 6b–i) and k-space images (Fig. 2c, d and Fig. 3c, f), by acquiring a background image, a reference image and a signal image. The background, reference and signal images were

recorded by acquiring the counts when no light was incident on the camera, when light was reflected from flat Ag on the sample and when light was reflected from the pattern of interest, respectively. The final reflectivity image was calculated using:

$$\begin{aligned}&\text{Reflectivity (\%)}\\&\quad = 100 \times (\text{Signal}-\text{Background})/(\text{Reference}-\text{Background})\end{aligned} \tag{7}$$

For the dispersed k-space measurements, a grating (150 lines mm^{-1} blazed at 500 nm) was inserted into the imaging path in the spectrometer such that the light was spectrally dispersed along one axis of the camera. The spectrometer slit was parallel to k_x. A linear polarizer was inserted into the collection path to select only p-polarized light, which couples to SPPs. Thus, in a single acquisition, the dispersion relation (energy versus in-plane momentum along the surface modulation, k_x, with $k_y \approx 0$) could be measured. The experimental window is overlaid with a schematic of the theoretical SPP dispersion in Extended Data Fig. 3b.

For the k-space images, a bandpass filter centred at 570 nm with a full-width at half-maximum (FWHM) of 10 nm was placed in the excitation path. The linear polarizer was removed from the detection path such that the measurement collected all polarizations equally. The slit at the entrance of the imaging spectrograph was opened completely and the k-space image was relayed to the camera using a mirror instead of a diffraction grating to eliminate stray diffracted light. A schematic of this measurement, performed at a narrow range of photon energies selected by the bandpass filter, is depicted in Extended Data Fig. 3d. A cartoon of the complete light cone and SPP dispersion is depicted in Extended Data Fig. 3c.

The reflectivity spectrum in Fig. 4d was obtained by plotting the dispersed k-space measurement for the three-component Fourier surface in Fig. 4b at a fixed angle of incidence (near normal incidence). Spectra were averaged over a collection angle of ±0.25°.

Analytical model

Optical modes bound to a periodic surface have an electric-field profile of the form

$$\mathbf{E}_\mathbf{k}(\mathbf{r}) = \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\mathbf{u}_\mathbf{k}(\mathbf{r}) \tag{8}$$

where $\mathbf{k}$ is the Bloch wavevector of the mode, and $\mathbf{u}_\mathbf{k}(\mathbf{r})$ is a function with the same periodicity as the surface. We consider a grating profile with modulation in one dimension, like those in Fig. 1, for which all surface Fourier components i have an in-plane wavevector $\mathbf{g}_i = g_i\hat{\mathbf{x}}$. The overall periodicity $2\pi/G$ of the surface profile can be much longer than any of the periodicities $\{2\pi/g_1, 2\pi/g_2, \ldots, 2\pi/g_N\}$ of the N constituent sinusoids:

$$G^{-1} = \mathrm{LCM}(g_1^{-1}, g_2^{-1}, \ldots, g_N^{-1}) \tag{9}$$

where LCM denotes the least common multiple. For example, the grating in Fig. 1g has an overall design periodicity of $2\pi/G = 96.6$ μm and $G = 0.0650$ μm^{-1}. The full field profile of a mode $\mathbf{E}_\mathbf{k}(\mathbf{r})$ contains all in-plane wavevector components $(k_x + nG,\ k_y)$ with any integer n. However, to calculate the plasmonic dispersion and stopbands of our Fourier surfaces in Fig. 1, we do not need the full field profile. Instead, we can use a relatively simple coupled-mode model with a limited basis, which only accounts for first-order coupling between plane waves differing in wavevector by $\mathbf{g}_i$ of one of the sinusoids of the grating.

On a flat Ag–dielectric interface, SPP modes have in-plane wavevector $\mathbf{k}_{\mathrm{SPP}}$ with magnitude:

$$k_{\mathrm{SPP}} = \frac{\omega}{c}\sqrt{\frac{\varepsilon_\mathrm{m}(\omega)\varepsilon_\mathrm{d}}{\varepsilon_\mathrm{m}(\omega)+\varepsilon_\mathrm{d}}} \tag{10}$$

where ω is the SPP angular frequency, c is the speed of light in vacuum and ε_m is the frequency-dependent relative permittivity of the metal. The relative permittivity of the dielectric ε_d is assumed to be frequency-independent. We note that when calculating $\mathbf{k}_{\mathrm{SPP}}$ for Figs. 2, 3, we used $\varepsilon_\mathrm{d} = 1.061$. This value was determined by fitting the SPP dispersion for an independent sample. Extracting a relative permittivity slightly above 1 was perhaps due to residual polymer on the Ag surface after templating. For the structures in Fig. 1, our fabrication process had been improved and $\varepsilon_\mathrm{d} = 1$ was extracted and used for modelling.

In Fig. 1, we measure the dispersion of our Fourier surfaces along the k_x direction. Stopbands in this direction arise whenever $2k_{\mathrm{SPP}} = g_i$ for one of the sinusoids i in the grating. This occurs at energies:

$$\hbar\omega_i = \frac{hc}{2n_{\mathrm{eff}}\Lambda_i} \tag{11}$$

where $\hbar = h/(2\pi)$ with h as Planck's constant, and $n_{\mathrm{eff}} = \sqrt{\varepsilon_\mathrm{m}(\omega)\varepsilon_\mathrm{d}/[\varepsilon_\mathrm{m}(\omega)+\varepsilon_\mathrm{d}]}$ is the effective refractive index of the SPP mode on the flat Ag–dielectric interface. Although the SPP dispersion, and any stopbands therein, lie outside the light cone, we can measure a stopband if some sinusoid j provides momentum to couple free-space photons to SPPs. The stopband will then appear in our reflectivity measurement at a photon in-plane wavevector with magnitude:

$$k_{ij} = n_{\mathrm{eff}}\frac{\omega_i}{c} - g_j = 2\pi\left(\frac{1}{2\Lambda_i} - \frac{1}{\Lambda_j}\right) \tag{12}$$

To calculate the stopbands and the SPP dispersion for our Fourier surfaces more rigorously, we use a coupled-mode model. We couple SPPs—surface waves with wavevector component $k_{x,0} = k_{\mathrm{SPP}}$—to surface waves with $k_{x,i} = k_{\mathrm{SPP}} - g_i$ for all sinusoids $i \in \{1, 2, \ldots, N\}$ in the surface profile. The coupling can be described by an interaction matrix H, which has dimensions $(N+1)\times(N+1)$. The diagonal elements of the matrix are the energies that a surface wave of wavevector component $k_{x,i}$ would have on a flat Ag–dielectric interface. We obtain these energies by evaluating the inverse of equation (10), $\omega(k_{\mathrm{SPP}})$, at $k_{\mathrm{SPP}} = |k_{x,i}|$:

$$H_{ii} = \hbar\omega(|k_{x,i}|) \tag{13}$$

For this, we use the relative permittivity data $\varepsilon_\mathrm{m}(\omega)$ of template-stripped Ag (ref. [43]) and $\varepsilon_\mathrm{d} = 1$ for air. The off-diagonal elements of the matrix, H_{ij}, describe the interaction between surface waves i and j. For simplicity, we consider only coupling involving the SPP wave, which has a wavevector component $k_{x,0} = k_{\mathrm{SPP}}$, and neglect coupling between surface waves with $i \geq 1$ and $j \geq 1$. Thus, the only non-zero off-diagonal elements of H are:

$$H_{0i} = H_{i0} = \hbar\Gamma_i \tag{14}$$

Here Γ_i is the (real-valued) rate at which the surface sinusoid i of the surface profile couples a surface wave with $k_{x,0}$ (that is, the SPP on a flat Ag–air interface) to a surface wave with $k_{x,i}$. This rate determines the width of the stopband $\Delta E_i \approx 2\hbar\Gamma_i$ owing to the grating component i. Extended Data Fig. 6 shows that we can control this by tuning the corresponding amplitude A_i of the sinusoid[30]. For Fig. 1i, we estimated values of Γ_i based on the dispersion data and plugged them into the model.

By solving for the eigenvalues of H, we obtain the energies E_i of the coupled modes. The eigenvectors $\mathbf{v}_i$ describe their composition in terms of the surface-wave basis functions. For each coupled mode, the first component of the eigenvector $v_{i,0}$ represents its SPP character.

So far, we have treated the coupling matrix H for a single value of k_{SPP}. However, to calculate dispersion plots such as those in Fig. 1, we must determine the eigenvalues and eigenvectors of H for a range of k_{SPP}. Thus, we considered a series of k_{SPP} values, labelled by $m \in \{1, 2, \ldots, M\}$, from 0 and 25 μm^{-1} in $M = 5{,}001$ steps of 0.005 μm^{-1}. At each $k_{\mathrm{SPP},m}$, the

corresponding coupling matrix H_m yields a set of $(N+1)$ mode energies $E_{m,i}$ and $(N+1)$ values $\upsilon_{m,i,0}$ for their corresponding SPP character (the coefficient for the contribution of the SPP with $k_{x,0} = k_{\mathrm{SPP},m}$ to the eigenvector of coupled mode j).

Now, in addition to coupling surface waves, we must include the fact that each sinusoid l in the surface profile ($l \in \{1, 2, \ldots, N\}$) can enable free-space photons to excite SPPs if $|k_{\mathrm{SPP}} - g_l| \leq \omega/c$ (that is, if the in-plane momentum required from the photon is inside the light cone). Free-space photons with in-plane wavevector $\mathbf{k}_{\parallel} = (k_x, 0)$ can excite SPPs if the momentum from sinusoid l matches coupled mode i with substantial SPP character $\upsilon_{i,0}$. We account for photon–SPP momentum-matching by considering the effect of grating components l again only in first order. Starting with the dispersion calculated in the last paragraph ($N+1$ energies $E_{m,i}$ at each $k_{\mathrm{SPP},m}$), we generate N copies of this dispersion by shifting the wavevector value to $k_{m,l} = k_{\mathrm{SPP},m} - g_l$ for all $l \in \{1, 2, \ldots, N\}$. These $k_{m,l}$ are the k values for which grating component l can in principle enable SPP incoupling. Then we copy and mirror the entire dispersion in the $(k = 0)$ axis, realizing that the entire problem is symmetric under inversion of the propagation direction of the modes. We thus obtain $2N$ copies of our calculated dispersion, some of which may fall entirely outside the experimental range of wavevectors and energies. We consider that at each point $(k_{m,l}, E_{m,i})$ or $(-k_{m,l}, E_{m,i})$, with $m \in \{1, 2, \ldots, M\}$, $i \in \{0, 1, 2, \ldots, N\}$ and $l \in \{0, 1, 2, \ldots, N\}$, the coupling to SPPs is proportional to $\Gamma_l \upsilon^2_{m,i,0}$. This reflects that, for first-order coupling, the magnitude of the admixture is proportional to the SPP character of the coupled mode. We thus obtain a model function for the incoupling V as a function of the photon in-plane wavevector component k_x and energy $\hbar\omega$ of:

$$V(k_x, \hbar\omega) = \sum_{m=1}^{M} \sum_{i=0}^{N} \sum_{l=0}^{N} \Gamma_l \upsilon^2_{m,i,0} \delta(k_x \pm k_{m,l}) \delta(\hbar\omega - E_{m,i}) \quad (15)$$

where δ is the Kronecker delta function. Finally, we broaden V by convolution with a function:

$$P(k_x) = \mathrm{sinc}^2(k_x d/2) \quad (16)$$

in the k_x direction that accounts for the finite length $d = 9$ µm of our gratings. We also convolute V with a Gaussian function $Q(\hbar\omega)$ with a variance of $\sigma^2 = (15\ \mathrm{meV})^2$ in the $\hbar\omega$ direction to match the experimental broadening. This arises from a combination of finite instrumental resolution, losses and the finite range of k_y values for reflected photons. The convolved function $(V * P * Q)(k_x, \hbar\omega)$ is plotted in Fig. 1c, f, i.

Quantification of diffraction efficiencies

We experimentally quantify the diffraction efficiencies of Fourier surfaces (Extended Data Fig. 8) with an optical k-space excitation and imaging setup. We illuminate the sample with monochromatic light at normal incidence and quantify the fraction of light that is diffracted and leaves the sample at off-normal angles. Light from a supercontinuum laser source (NKT, Fianium, repetition rate 7.8 MHz) was filtered to a linewidth of about 1 nm using a tunable filter box (NKT, LLTF Contrast) and was collimated after the output of a single-mode fibre using an objective (Nikon, TU Plan Fluor 10×, NA 0.3). After passing through a 750-nm short-pass filter, a fraction of the beam was directed to a power meter using a beam splitter. The remaining beam was sent through a reflective neutral-density filter and a linear polarizer (polarization direction, s or p, as specified in Extended Data Fig. 8) before being focused onto the centre of the back focal plane of a microscope objective (Nikon, TU Plan Fluor 50 × , NA 0.8) using a lens with focal length $f = 750$ nm (placed a distance f before the back focal plane). In this optical configuration the sample of interest in the focus of the microscope objective was illuminated with light from a narrow set of solid angles centred around normal incidence. The finite size of the focused laser beam on the back focal plane resulted in a defocused Gaussian illumination spot on the investigated sample. The light reflected and diffracted by the sample was collected through the same microscope objective, redirected with a beam splitter and used to image the back focal plane of the microscope objective onto a sensitive digital camera (Andor, Zyla PLUS sCMOS). A real-space aperture in the relay system of the collection path was reduced to a diameter comparable to the side length of the Fourier surface. In this optical configuration, the illumination wavelength λ was varied between 450 nm and 700 nm in steps of 1 nm while recording one back focal plane image per wavelength step with 5 ms acquisition time. This process was done subsequently for the Fourier surface under investigation and for flat Ag as a reference. A separate image without laser illumination was subtracted from each k-space image to remove the background counts of the detector. The k-space images were subsequently corrected for power fluctuations of the supercontinuum source (as measured with the power meter), resulting in two sets of k-space images for the investigated Fourier surface and the flat Ag reference, respectively. The k-space images of the reference sample showed a bright spot centred around $k_x = k_y = 0$ with summed intensity $I_{\mathrm{ref}}(\lambda)$, corresponding to specular reflection of the beam impinging on and exiting from the flat reference surface at normal incidence. For Fourier surfaces periodic along the x direction additional spots were observed centred around $k_x = g_i$, $k_y = 0$ with $i = \pm 1$. The intensity of each spot $I_i(\lambda)$ was extracted by summing the corresponding pixels of the k-space images. The diffraction efficiencies were calculated as $\eta_i(\lambda) = I_i(\lambda)/I_{\mathrm{ref}}(\lambda)$, corresponding to the fraction of impinging photons diffracted into diffraction order i. We note that this formula neglects reflection losses from flat Ag (a few per cent).

Data availability

The data supporting the findings of this study are available from the corresponding author on reasonable request.

42. Moreno, V., Román, J. F. & Salgueiro, J. R. High efficiency diffractive lenses: deduction of kinoform profile. *Am. J. Phys.* **65**, 556–562 (1997).
43. McPeak, K. M. et al. Plasmonic films can easily be better: rules and recipes. *ACS Photonics* **2**, 326–333 (2015).
44. Kapsalidis, F. et al. Dual-wavelength DFB quantum cascade lasers: sources for multi-species trace gas spectroscopy. *Appl. Phys. B* **124**, 107 (2018).

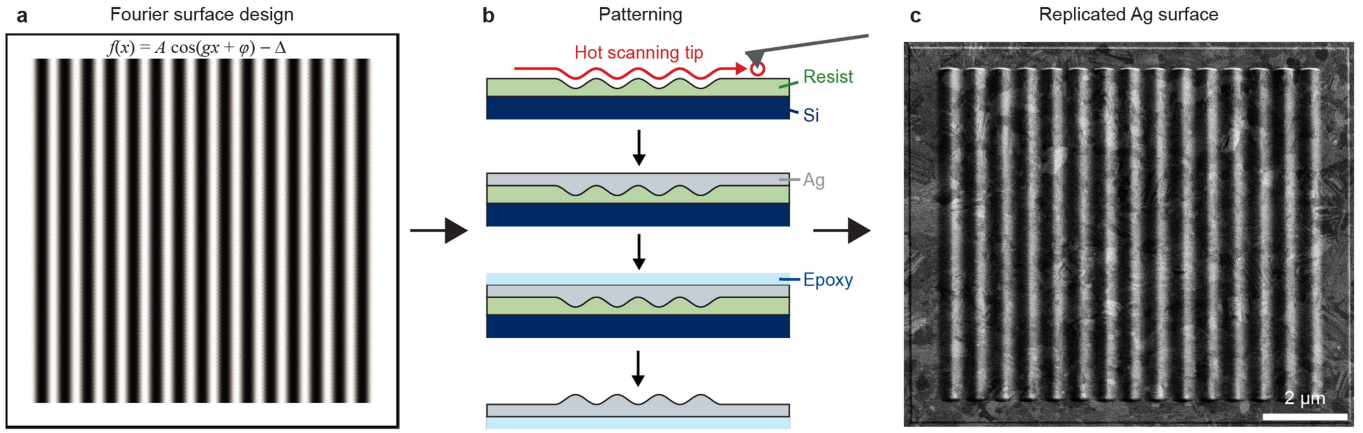
Acknowledgements We thank S. Bonanni, U. Drechsler, F. Enz, T. Kulmala, A. Olziersky and R. Stutz for technical assistance and D. Chelladurai, U. Dürig, R. Keitel, A. Knoll, M. Kohli, N. Rotenberg, D. Thureja, J. Winkler and H. Wolf for discussions. This project was funded by the European Research Council under the European Union's Seventh Framework Program (FP/2007-2013)/ERC Grant Agreement Number 339905 (QuaDoPS Advanced Grant). F.T.R. (Rubicon-680-50-1509, Gravitation Program "Multiscale Catalytic Energy Conversion", Veni-722.017.002), S.J.W.V. (OCENW.KLEIN.008) and B.l.F. (Rubicon-680-50-1513) acknowledge support from the Netherlands Organisation for Scientific Research.

Author contributions N.L., B.l.F. and D.J.N. conceived the project. N.L., R.B. and S.J.W.V. designed the Fourier surfaces with input from K.R., F.T.R. and D.J.N. N.L. patterned the polymer surfaces with assistance from K.R., M.S. and S.B. N.L. and R.B. transferred the patterns to optical materials with assistance from K.R. and M.S. N.L. performed the characterization and topography analysis of the Fourier surface structures. N.L. and R.B. performed the optical experiments. N.L., R.B., S.J.W.V. and F.T.R. analysed the optical data. F.T.R. developed the analytical model. N.L. and D.J.N. wrote the manuscript with input from all authors. D.J.N. supervised the project.

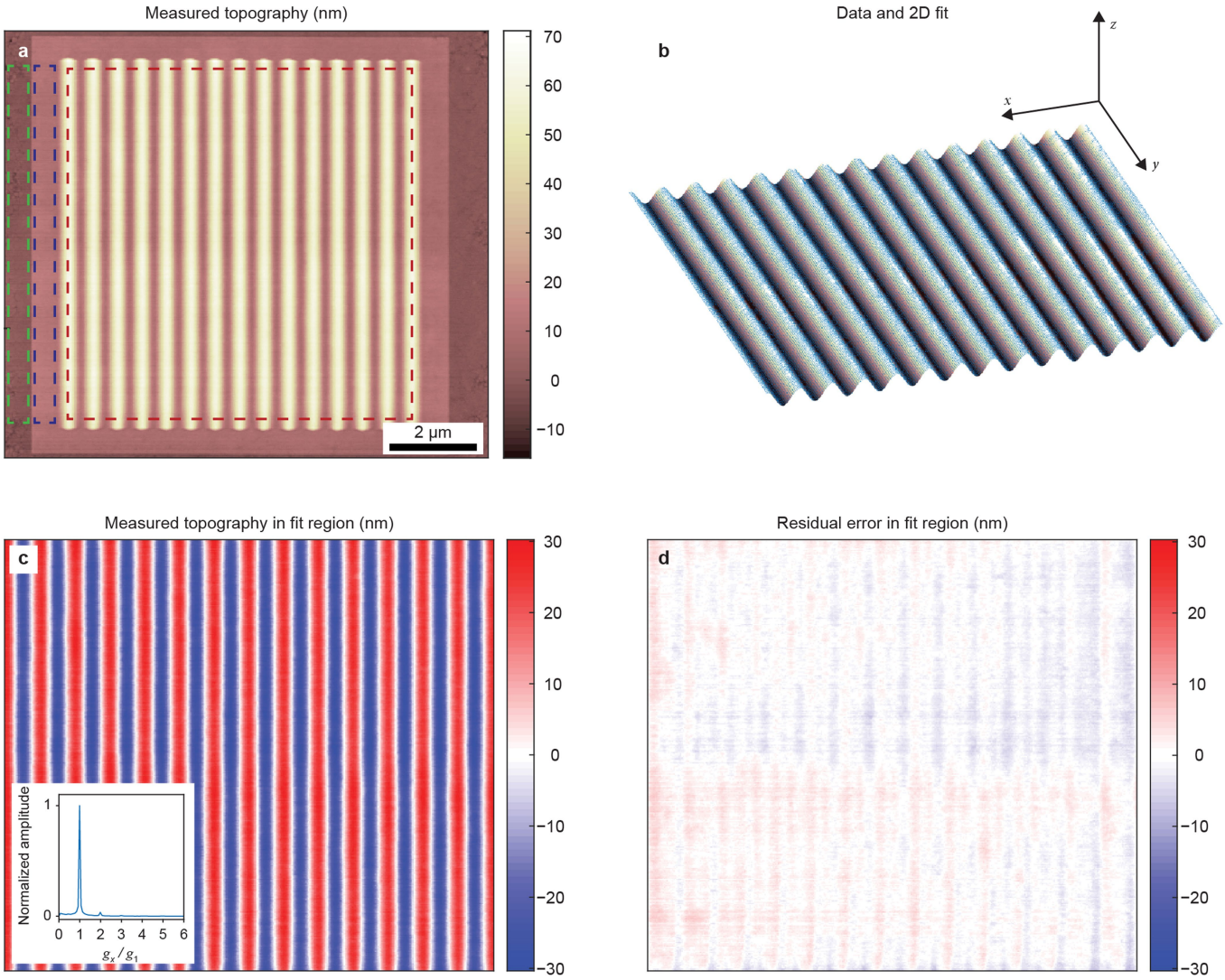
Competing interests The authors declare the following potential competing financial interests: S.B. is employed by Heidelberg Instruments Nano (previously SwissLitho AG), a provider of thermal scanning-probe lithography tools. At the time of his contribution, M.S. worked for SwissLitho AG. N.L., R.B., F.T.R. and D.J.N. have filed a patent application related to ideas in this work.

Additional information
Correspondence and requests for materials should be addressed to D.J.N.

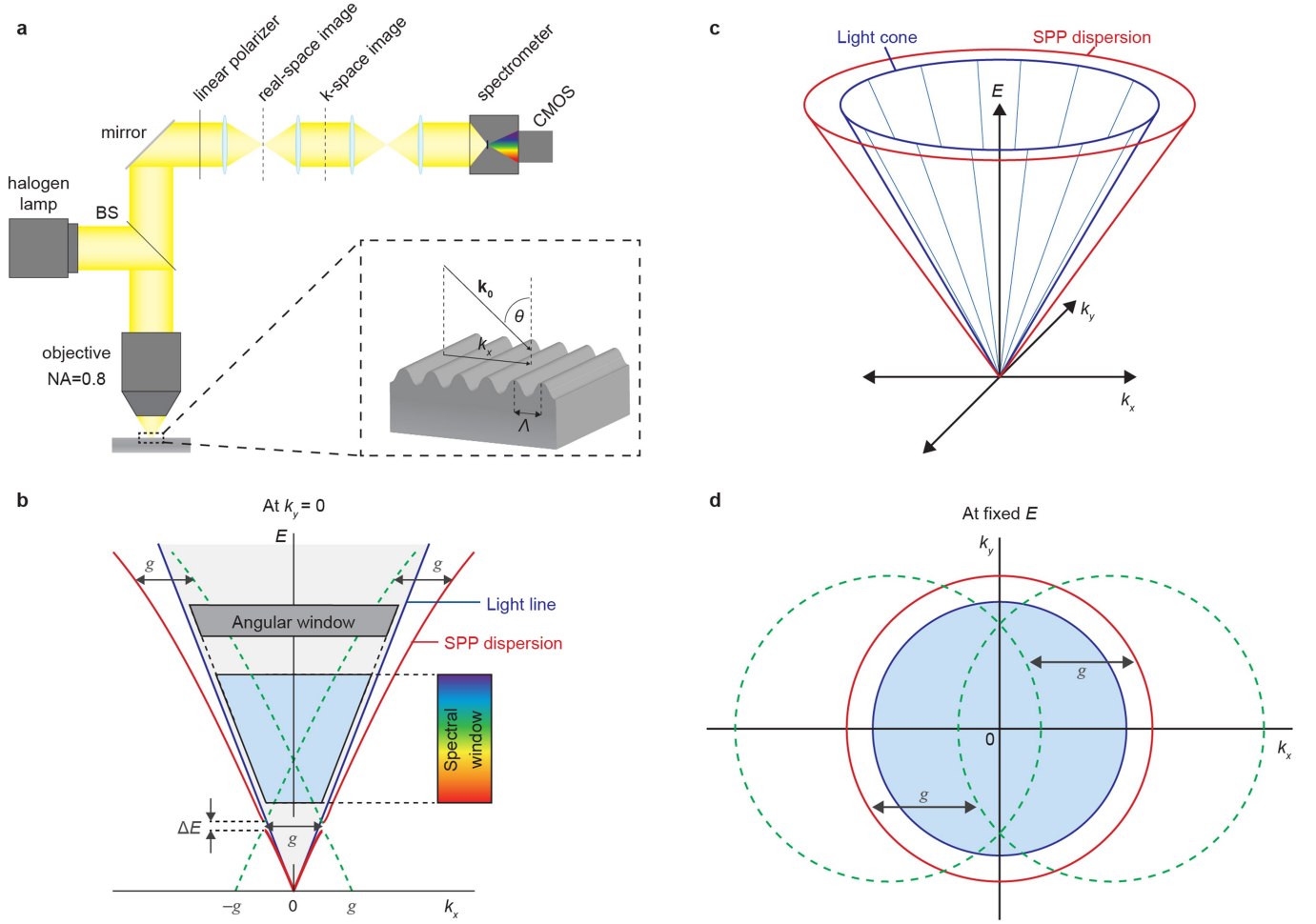
Peer review information *Nature* thanks Wei-Ting Chen, Maryna Meretska and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



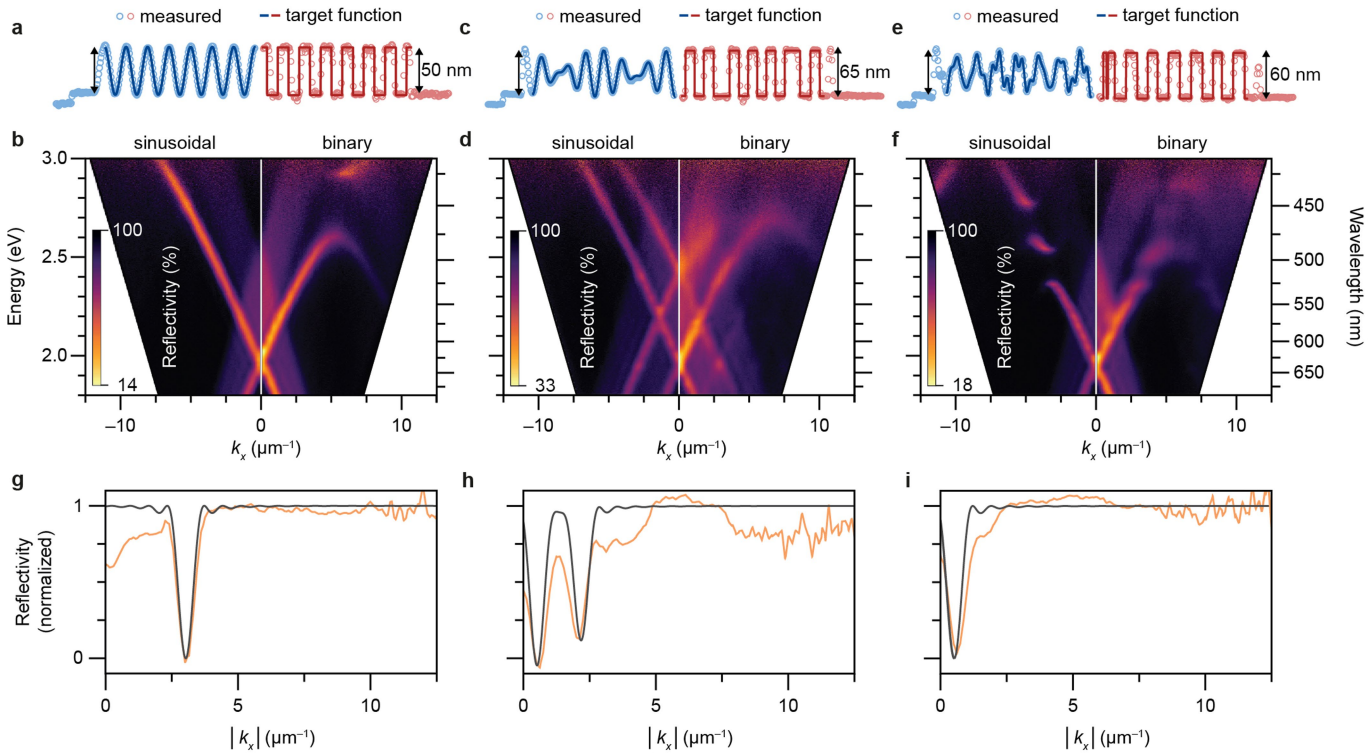
Extended Data Fig. 1 | Design and fabrication of Fourier surfaces. a, Design of a Fourier surface. The analytical formula for the desired surface profile (here, a single sinusoid modulated in 1D) is converted into a grayscale bitmap. Each 10 nm × 10 nm pixel has a depth level between 0 and 255 (8-bit). The bitmap contains the sinusoidal function in the horizontal direction within the white border, which is constant along the vertical direction. The pixels in the white border are set to the minimum depth level. **b**, Process flow showing the patterning steps for Ag Fourier surfaces: (i) The hot scanning tip is used to create a single sinusoid in the polymer resist, (ii) an optically thick (>500 nm) Ag layer is thermally evaporated onto the polymer, (iii) a glass microscope slide is affixed to the back of the Ag layer using ultraviolet-curable epoxy, and (iv) the glass/epoxy/Ag stack is stripped off the polymer film. Alternative fabrication pathways for transferring the Fourier surface pattern to other materials are presented in the Methods. **c**, SEM (30° tilt) of a single-component Fourier surface transferred to Ag via templating. The initial analytical design is replicated accurately in the final Ag surface.



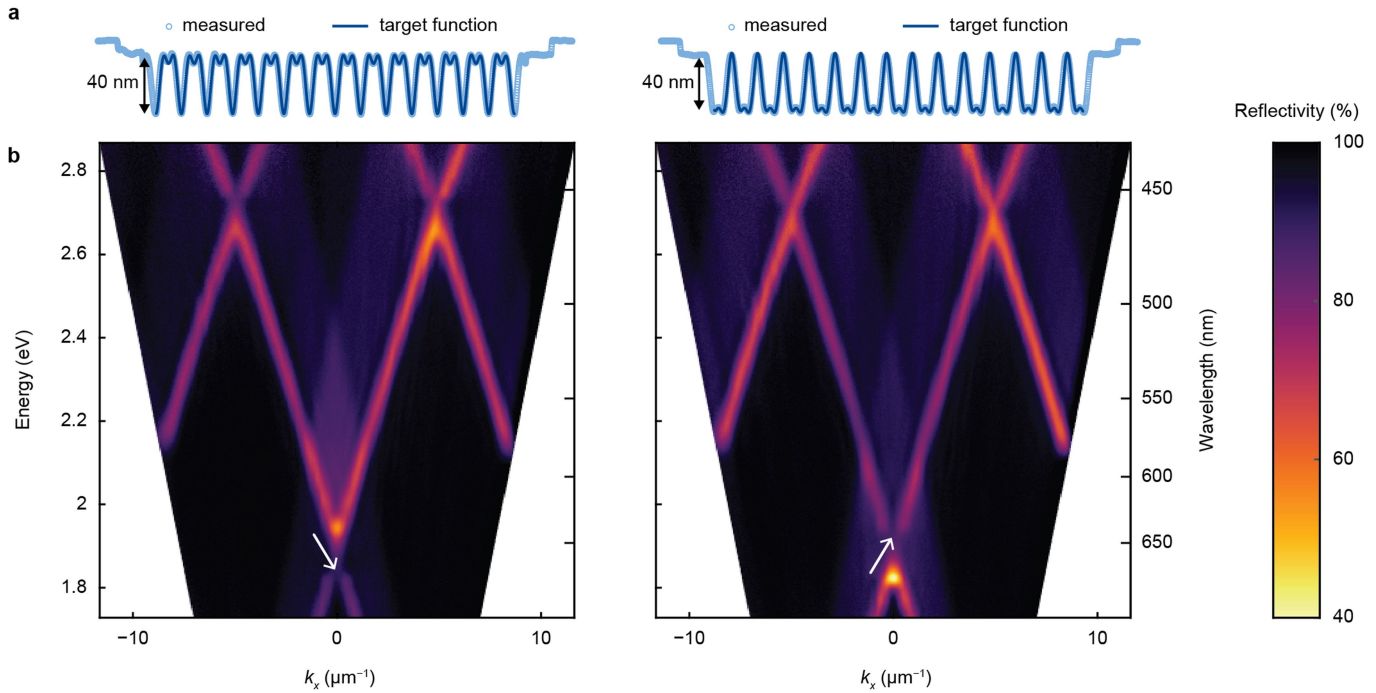
Extended Data Fig. 2 | Topography characterization. a, AFM micrograph of the measured topography (colour scale) for a single-sinusoidal Ag grating. The RMS roughness of the unpatterned flat Ag film is 1.6 nm, extracted from the area indicated by the green dashed box. The RMS roughness of the patterned flat Ag film is 1.3 nm, extracted from the area indicated by the blue dashed box. The area indicated by the red dashed box is used for fitting and analysis of the surface profile. **b**, 2D fit of a sinusoidal function (yellow/brown surface) to topography data (blue dots) from the region indicated in the red dashed box in **a**. The amplitude of the fitted function is $A_1 = 25.5$ nm (2% larger than design value) with a period of $\Lambda = 610$ nm (1.7% larger than design value). Such horizontal errors were consistent over many samples and attributed to a distance miscalibration in the thermal scanning probe. The RMS error between the design function and measured topography was found to be 1.8 nm after this horizontal error was taken into account. **c**, Measured topography (colour scale) of the structure in **a**, plotted only for the fit region (red dashed box in **a**), scaled from the minimum depth value to the maximum depth value and centred at zero. The inset shows a line cut (along g_x at $g_y = 0$, where g_x and g_y are the components of **g** along the x and y axes, respectively) from the 2D Fourier transform of the measured topography in the fit region, normalized to the peak value at g_1. The second harmonic at $g_x/g_1 = 2$ is barely visible and has an amplitude of 3.5% of the peak at $g_x/g_1 = 1$, corresponding to a real-space amplitude of 0.9 nm. **d**, Residual error (colour scale) between the data and the fitted function, plotted for the fit region as in **c**. For comparison, the data are scaled over the same range as in **c**, centred at 0.



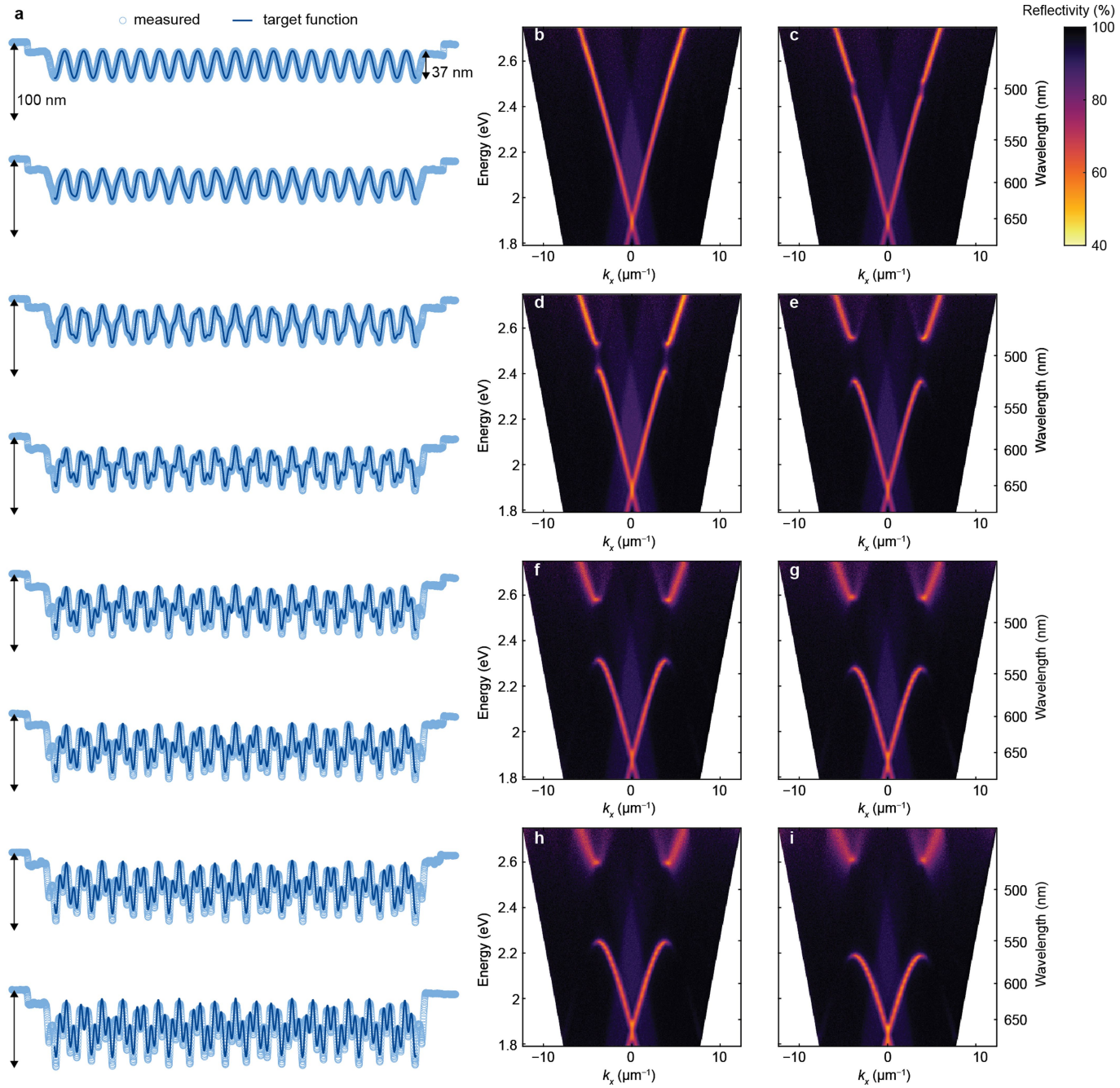
Extended Data Fig. 3 | Optical measurement of plasmonic Fourier surfaces. **a**, Schematic of the optical setup used for *k*-space reflectivity measurements. Further details are in the Methods. The inset shows a vector diagram of light with wavevector $\mathbf{k}_0$ incident at angle θ on a Fourier surface pattern with period Λ. BS, beamsplitter. CMOS, complementary metal-oxide-semiconductor digital camera. **b**, Schematic of the dispersion diagram (energy versus in-plane wavevector component, k_x) for free-space photons incident on a sinusoidal grating with $k_y = 0$ (as in Fig. 1). By tuning θ, photons have access to the shaded region inside the light lines (solid blue lines). The red lines show the SPP dispersion, k_{SPP}. Dashed green curves indicate the SPP dispersion displaced by the grating spatial frequency g. Inside the light line, these curves represent where free-space photons can couple to SPPs, and vice versa (that is, where $k_x \pm g = k_{SPP}$). A stopband of width ΔE opens when counter-propagating SPPs are coupled by g. The blue trapezoidal region depicts the experimentally accessible area on the dispersion diagram, limited by the spectral window of the spectrometer along E, and the angular window of reflected light collected by the microscope objective along k_x. **c**, Schematic of the dispersion diagram for free-space photons incident on a surface, plotted for both in-plane wavevectors, k_x and k_y. The light line and SPP dispersion in **b** are both cones (blue and red lines, respectively). **d**, A slice through the dispersion diagram in **c** at fixed energy. Free-space photons incident on a surface can have wavevectors inside the light cone (blue-shaded region). The SPP dispersion is the larger red circle. Dashed green circles show solutions to $\mathbf{k}_{\parallel} \pm \mathbf{g} = \mathbf{k}_{SPP}$. In this example, $\mathbf{g} = g\hat{\mathbf{x}}$.



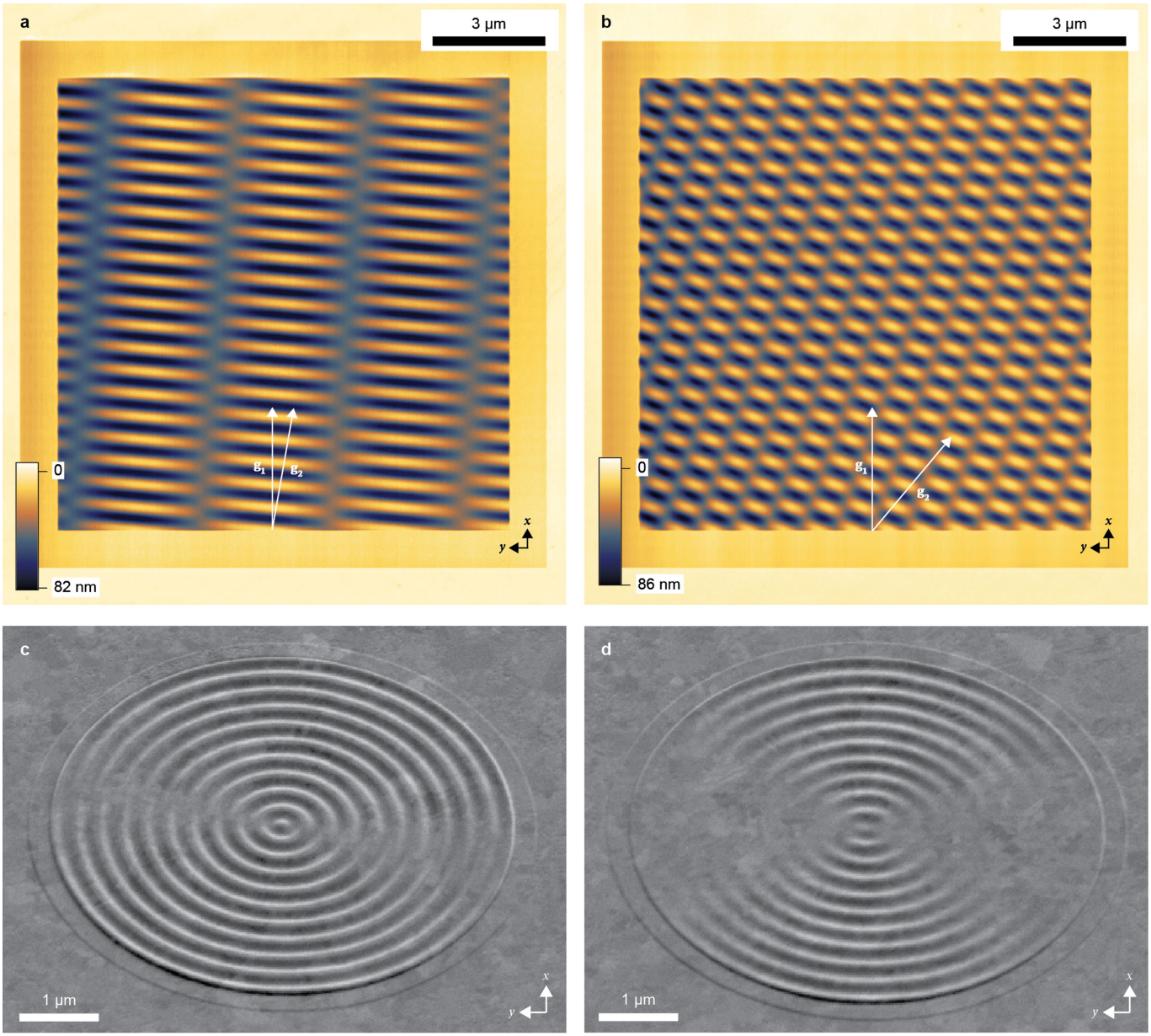
Extended Data Fig. 4 | Further analysis of the Ag Fourier surfaces in Fig. 1. **a**, **c**, **e**, Measured (AFM) and targeted surface topographies for the same sinusoidal structures as in Fig. 1a, d and g (blue) and their 'binarized' versions (red). The data for the Fourier surfaces represent half of the scans shown in Fig. 1b, e, h. The scan lengths for the binarized versions are 5.9 μm, 6.0 μm and 5.5 μm, respectively. For each Fourier surface, a binarized profile was obtained using a published thresholding procedure (see Methods). These binarized structures were then fabricated in a Si substrate using electron-beam lithography and etching. Ag replicas were obtained by templating (see Methods). The depth scale bars are 50 nm, 65 nm and 60 nm for both structures in **a**, **c** and **e**, respectively. **b**, **d**, **f**, Comparison of experimental angle-resolved reflectivity spectra measured for the sinusoidal surfaces shown in **a**, **c** and **e** (left) and their binarized versions (right). The data for the sinusoids are the same as the left sides of Fig. 1c, f, i. The optical responses of the binarized gratings are clearly corrupted by the unwanted spatial frequencies in the structure. **g**–**i**, Normalized line cuts (orange curves) through the reflectivity data shown in the left panels in **b**, **d** and **f** at 500 nm, 600 nm and 600 nm, respectively. The black curves show the predicted reflectivity versus the absolute value of the in-plane wavevector, $|k_x|$ from our model (see Methods). The comparison reveals good agreement between the model and the data without any adjustable parameters (other than the normalization). For all structural design parameters, see Extended Data Table 1.



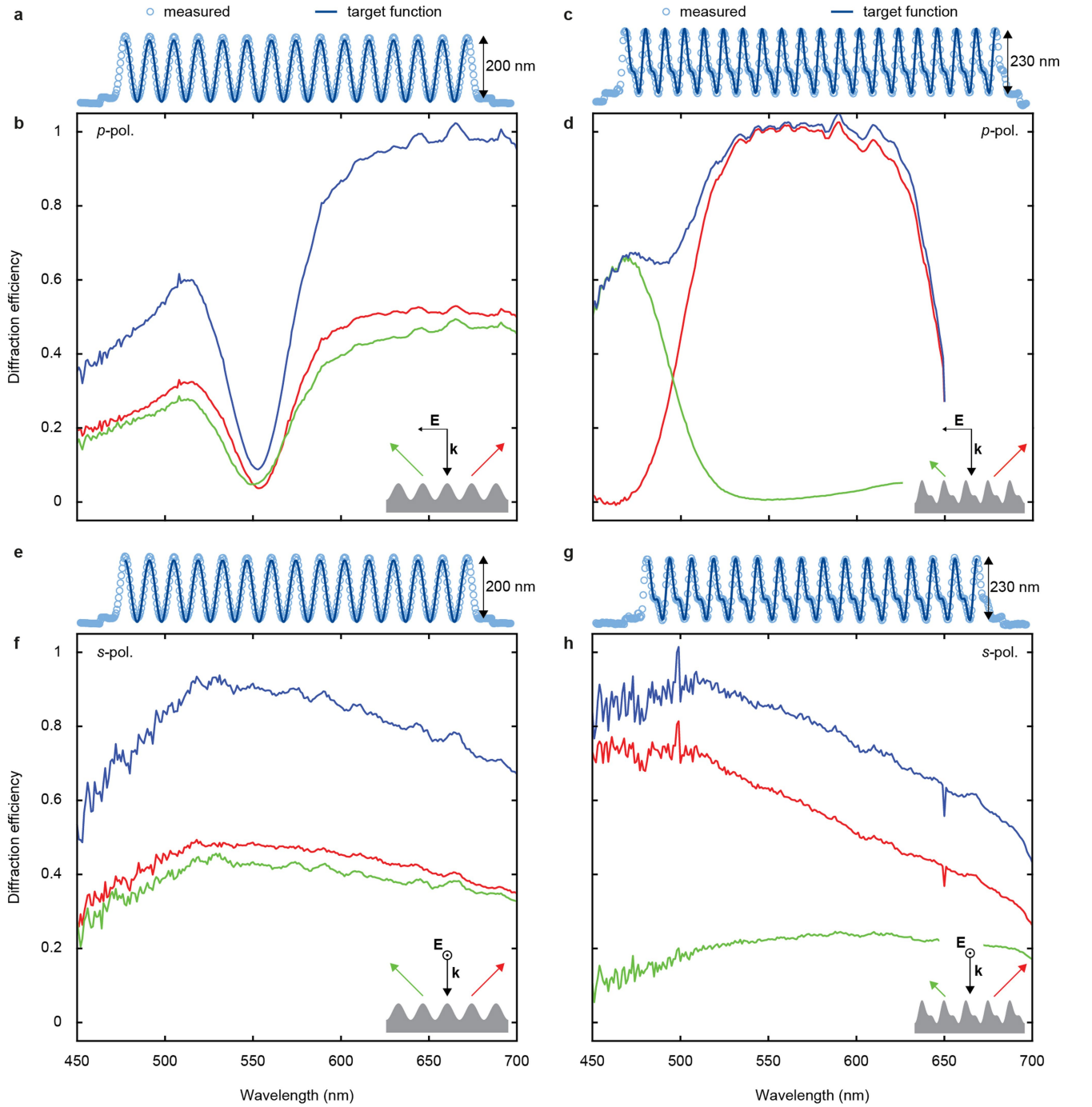
Extended Data Fig. 5 | Control of 'dark' band edges in two-component sinusoidal gratings. a, Comparisons of the measured (light blue points) and targeted surface topographies (dark blue lines) in the polymer surface, measured during patterning. Scan lengths are 11.5 μm. The left grating has the height profile $f(x) = A_1\cos(gx + \pi) + (A_1/2)\cos(2gx + \varphi_2) - \Delta$ with $\varphi_2 = \pi$. The grating on the right has the same $f(x)$ except $\varphi_2 = 0$. **b**, The measured reflectivity in k-space (as in Fig. 1) for Ag gratings templated from the structures in **a**. In both the left and right gratings, a stopband opens near 1.9 eV, but the choice of phase can control whether an optically dark state exists at the lower (left) or upper (right) band edge. The band edge with the optically dark state is marked with white arrows. For all structural design parameters, see Extended Data Table 1.



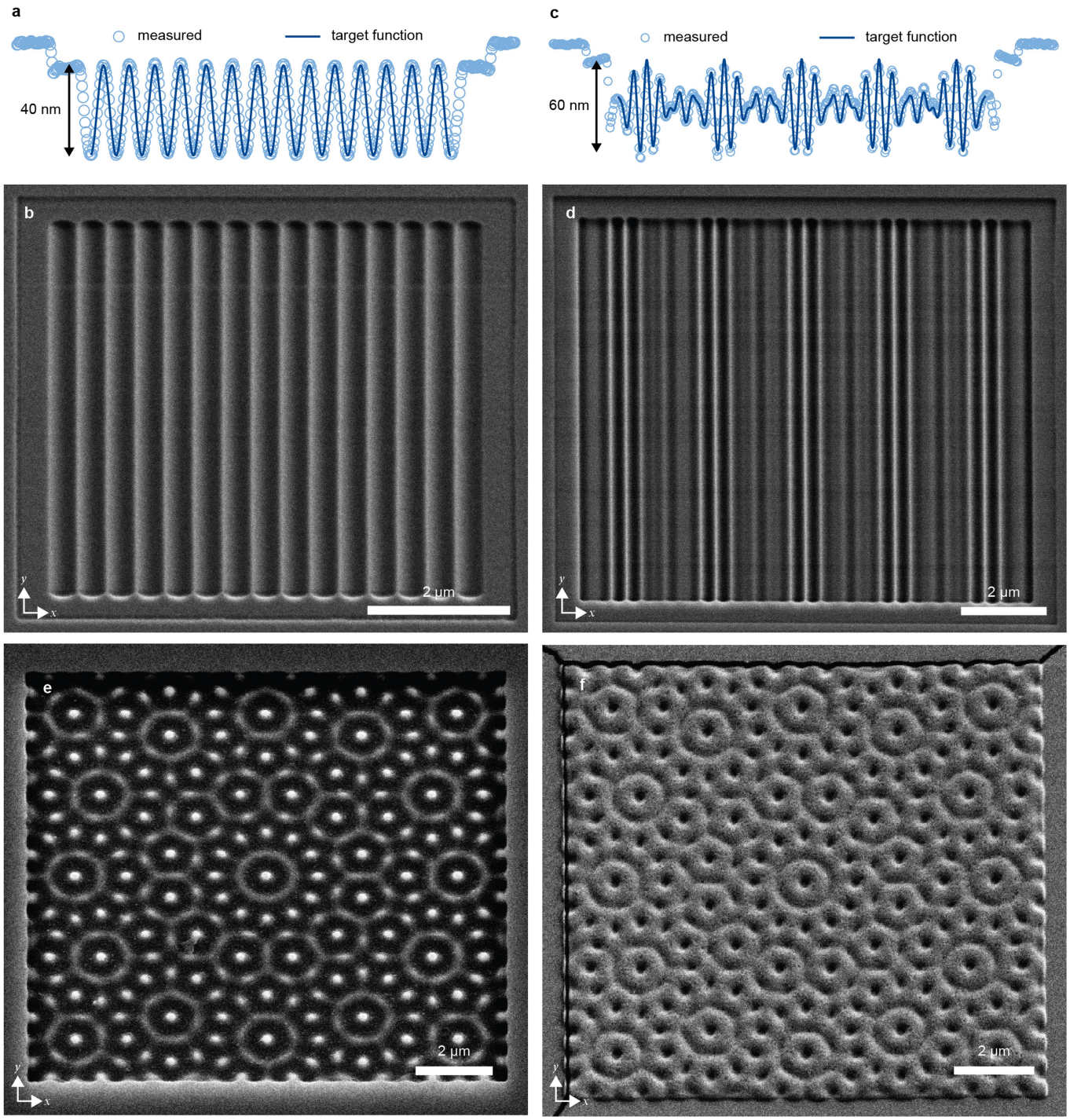
Extended Data Fig. 6 | Control of stopband width in two-component sinusoidal gratings. a, Comparisons of the measured (light blue points) and targeted surface topographies (dark blue lines) in the polymer surfaces, measured during patterning, for structures exhibiting a single stopband. Scan lengths are 14.5 μm and the left vertical scale bar is 100 nm for all scans. From top to bottom: a series of two-component sinusoidal gratings, where $A_1 = 18.5$ nm, $\Lambda_1 = 620$ nm, A_2 is varied and $\Lambda_2 = 230$ nm. A_2 has values of 0 nm, 2.5 nm, 5 nm, 10 nm, 15 nm, 18 nm, 20 nm and 25.1 nm. **b**–**i**, Measured plasmonic dispersion diagrams for Ag gratings templated from the profiles in **a**, from top to bottom, respectively. The width of the stopband increases because A_2 is the amplitude of the Fourier component responsible for creating the plasmonic stopband. For all structural design parameters, see Extended Data Table 1.



Extended Data Fig. 7 | Fourier surface patterns. a, Measured topography (obtained during patterning) of the polymer film (PMMA/MA; see Methods) used to template the structure in Fig. 2a. The two spatial-frequency vectors $\mathbf{g}_1$ and $\mathbf{g}_2$ that define the surface profile are overlaid on the pattern. Here, $\mathbf{g}_1$ and $\mathbf{g}_2$ have the same magnitude $g_1 = g_2 = 2\pi/600$ nm, and $\mathbf{g}_2$ is rotated −10° from $\mathbf{g}_1$, where $\mathbf{g}_1$ lies along $\hat{\mathbf{x}}$. **b**, As in **a**, but the template corresponding to the structure in Fig. 2b. Again, $\mathbf{g}_1$ and $\mathbf{g}_2$ have the same magnitude $g_1 = g_2 = 2\pi/600$ nm, but $\mathbf{g}_2$ is rotated −40° from $\mathbf{g}_1$, where $\mathbf{g}_1$ lies along $\hat{\mathbf{x}}$. **c**, SEM (45° tilt) of a circular sinusoidal Ag grating with $\Lambda = 600$ nm. **d**, SEM (45° tilt) of two superimposed circular sinusoidal gratings, as in **c**, each with $\Lambda = 600$ nm. The centre of one grating is translated +150 nm and the other −150 nm in $\hat{\mathbf{y}}$ from the origin in the middle of the pattern. The spatial interference results in a moiré pattern with broken circular symmetry. For all structural design parameters, see Extended Data Table 1.



Extended Data Fig. 8 | Efficient diffraction from photonic Fourier surfaces. **a**, Comparison of the measured (AFM) and targeted surface topography (accounting for a slight distance miscalibration and depth nonlinearity in the thermal scanning probe) for a deeper sinusoidal Ag grating designed for efficient optical diffraction. The scan length is 18.5 µm. **b**, Measured diffraction efficiency as a function of photon wavelength for the grating profile shown in **a**, for p-polarized illumination at normal incidence. The inset illustrates how incident photons (black arrow with wavevector **k** and electric field **E** indicated) diffract symmetrically into the +1 (red arrow) and −1 (green arrow) diffraction orders. The red and green curves correspond to the +1 and −1 diffracted intensities, respectively, normalized to the intensity reflected from an unpatterned flat reference spot on the same Ag sample. The sum of the red and green curves (blue line) peaks at about 97%. Owing to fluctuations in the collected diffraction intensity, the measured efficiencies have an estimated error of ±5%. We also note that our measurement does not account for reflection losses in the Ag (a few per cent). **c**, As in **a**, but for a two-component sinusoidal Ag grating where the relative phase between the two components is chosen to break the mirror symmetry of the structure about the *y*–*z* plane. **d**, As in **b**, but now the broken symmetry causes nearly all of the incident light to be diffracted into the +1 diffraction order (red curve) for a given wavelength range. **e**–**h**, As in **a**–**d**, but for s-polarized illumination at normal incidence. For all structural design parameters, see Extended Data Table 1.



Extended Data Fig. 9 | SiN_x and TiO_2 Fourier surfaces. a, Comparison of the measured (AFM) and targeted surface topography (accounting for a slight distance miscalibration in the thermal scanning probe) for a single sinusoid in SiN_x, transferred via reactive-ion etching (see Methods). Scan length is 11.3 µm. **b**, SEM (30° tilt) of the same structure in **a**. The final profile in SiN_x has a measured RMS error of 2.5 nm using the same methodology as in Extended Data Fig. 2. **c**, As in **a**, but for a three-component SiN_x grating. Scan length is 14.8 µm. **d**, As in **b**, but for the structure in **c**. The final profile in SiN_x has a measured RMS error of 3.9 nm using the same methodology as in Extended Data Fig. 2. **e**, SEM (30° tilt) of a 12-fold rotationally symmetric quasicrystal, as in Fig. 3d, transferred from the patterned polymer into Si via inductively coupled plasma etching (see Methods). **f**, SEM (30° tilt) of the pattern in **e** transferred into a TiO_2 thin film via template stripping (see Methods). For all structural design parameters, see Extended Data Table 1.

Extended Data Table 1 | Design parameters for Fourier surfaces

	Parameters										Height profile	
Figure	A_1 (nm)	A_2 (nm)	A_3 (nm)	Λ_1 (nm)	Λ_2 (nm)	Λ_3 (nm)	φ_1 (deg)	φ_2 (deg)	φ_3 (deg)	Δ (nm)	$f(x)$	
Fig. 1a	25.0	–	–	600	–	–	180	–	–	35.0	$\sum_i A_i \cos(g_i x + \varphi_i) - \Delta$	
Fig. 1d	18.1	14.3	–	600	475	–	0	0	–	42.3		
Fig. 1g	18.4	7.0	6.4	600	230	210	0	0	0	41.8		
ED Fig. 5	19.3	9.6	–	620	310	–	180	180	–	24.4		
								0		38.9		
ED Fig. 6	18.5	0 to 25.1	–	620	230	–	0	0	–	28.0 to 53.1		
Fig. 4b	12.0	12.0	12.0	620	520	445	0	0	0	45.9		
ED Fig.8a,e	100	-	-	1000	-	-	180	-	-	110		
ED Fig.8c	88.6	44.3	-	800	400	-	90	90	-	125		
ED Fig.8g	88.6	44.3	-	900	450	-	90	90	-	125		
ED Fig.9b	25.0	–	–	400	–	–	180	–	–	35.0		
ED Fig.9d	14.1	11.8	10.1	414	347	297	0	0	0	45.8		

Figure	i	A_i (nm)	Λ_i (nm)	φ_i (deg)	$\theta_{1,7}$ (deg)	$\theta_{2,8}$ (deg)	$\theta_{3,9}$ (deg)	$\theta_{4,10}$ (deg)	$\theta_{5,11}$ (deg)	$\theta_{6,12}$ (deg)	Δ (nm)	$f(x,y)$
Fig. 2a	1,2	17.5	600	0	0	-10	–	–	–	–	45.0	$\sum_i A_i \cos[g_i(x\cos\theta_i + y\sin\theta_i) + \varphi_i] - \Delta$
Fig. 2b	1,2	17.5	600	0	0	-40	–	–	–	–	45.0	
Fig. 3a	1,2,3	15.6	600	0	0	60	120	–	–	–	56.7	
Fig. 3d	1–6	10.0	600	0	0	30	60	90	120	150	70.0	
Fig. 4e	1–6	5.6	700	0	0	30	60	90	120	150	77.3	
	7–12	5.6	308	0	0	30	60	90	120	150	77.3	
ED Fig. 9e,f	1–6	11.1	615	0	0	30	60	90	120	150	76.7	

Figure	i	A_1 (nm)	A_2 (nm)	Λ_1 (nm)	Λ_2 (nm)	φ_1 (deg)	φ_2 (deg)	Δ (nm)	$\mathbf{r_1}$ (nm)	$\mathbf{r_2}$ (nm)	$f(r,\theta)$
ED Fig. 7c	1	35.0	–	600	–	180	–	45	$\mathbf{0}$	–	$\sum_i A_i \cos(g_i\lvert\mathbf{r} - \mathbf{r}_i\rvert + \varphi_i) - \Delta$
ED Fig. 7d	1,2	17.5	17.5	600	600	180	0	45	$-150\,\hat{\mathbf{y}}$	$150\,\hat{\mathbf{y}}$	

Figure	A (nm)	L (nm)	Δ (nm)	$f(r)$
Fig. 2e	35.0	1581	45.0	$A\sin\left[\pi\left(\frac{r}{L}\right)^2\right] - \Delta$

Design parameters for all Fourier surfaces demonstrated in this work. The functions are defined for the design to be patterned in the polymer surface, where x and y lie in-plane and z is perpendicular (pointing away from the substrate). A right-handed coordinate system is used with the origin placed in the middle of the pattern in both the x and y directions. In these formulas, the height of the surface is defined relative to the unpatterned flat surface where $z = 0$. All A_i and Δ_i (Λ_i) have been rounded to the nearest 0.1 nm (1.0 nm). Analysis of the measured topographies for templated Ag gratings shows that the Λ_i values are consistently about 2% larger than the design value (Extended Data Fig. 2), attributed to a distance miscalibration in the thermal scanning probe. See Methods.

Article

Ultralow-dielectric-constant amorphous boron nitride

https://doi.org/10.1038/s41586-020-2375-9

Received: 16 November 2019

Accepted: 25 March 2020

Published online: 24 June 2020

Check for updates

Seokmo Hong[1], Chang-Seok Lee[2], Min-Hyun Lee[2], Yeongdong Lee[3,4], Kyung Yeol Ma[4,5], Gwangwoo Kim[1], Seong In Yoon[4,5], Kyuwook Ihm[6], Ki-Jeong Kim[6], Tae Joo Shin[1,7], Sang Won Kim[2], Eun-chae Jeon[8], Hansol Jeon[3], Ju-Young Kim[3], Hyung-Ik Lee[9], Zonghoon Lee[3,4], Aleandro Antidormi[10], Stephan Roche[10,11], Manish Chhowalla[12]✉, Hyeon-Jin Shin[2]✉ & Hyeon Suk Shin[1,4,5,13]✉

Decrease in processing speed due to increased resistance and capacitance delay is a major obstacle for the down-scaling of electronics[1–3]. Minimizing the dimensions of interconnects (metal wires that connect different electronic components on a chip) is crucial for the miniaturization of devices. Interconnects are isolated from each other by non-conducting (dielectric) layers. So far, research has mostly focused on decreasing the resistance of scaled interconnects because integration of dielectrics using low-temperature deposition processes compatible with complementary metal–oxide–semiconductors is technically challenging. Interconnect isolation materials must have low relative dielectric constants (κ values), serve as diffusion barriers against the migration of metal into semiconductors, and be thermally, chemically and mechanically stable. Specifically, the International Roadmap for Devices and Systems recommends[4] the development of dielectrics with κ values of less than 2 by 2028. Existing low-κ materials (such as silicon oxide derivatives, organic compounds and aerogels) have κ values greater than 2 and poor thermo-mechanical properties[5]. Here we report three-nanometre-thick amorphous boron nitride films with ultralow κ values of 1.78 and 1.16 (close to that of air, $\kappa = 1$) at operation frequencies of 100 kilohertz and 1 megahertz, respectively. The films are mechanically and electrically robust, with a breakdown strength of 7.3 megavolts per centimetre, which exceeds requirements. Cross-sectional imaging reveals that amorphous boron nitride prevents the diffusion of cobalt atoms into silicon under very harsh conditions, in contrast to reference barriers. Our results demonstrate that amorphous boron nitride has excellent low-κ dielectric characteristics for high-performance electronics.

Modern high-performance logic and memory devices used in multifunctional electronics are constructed using materials and designs that have enabled a drastic reduction of transistor size and the packing of more circuits in smaller areas[1–3,6–11]. However, the reduction in the dimensions of metal interconnects and the increased packing density have led to an increase in the resistance (R) and capacitance (C) delay, which is becoming comparable to the operation speed of the devices. Ideally, both R and C should be simultaneously reduced to achieve continuous scaling of devices. However, the development of electrically, mechanically and thermally robust low-κ materials ($\kappa < 2$) using complementary metal–oxide–semiconductor (CMOS)-compatible processes that are good inter-metal and inter-layer dielectrics and act as diffusion barriers against electro-migration of metal atoms from interconnects has been challenging.

State-of-the-art strategies for achieving low-κ dielectrics have involved reducing the polarization strength and density of SiO_2 ($\kappa = 4$) by incorporating fluorine ($\kappa = 3.7$ for SiOF) or CH_3 ($\kappa = 2.8$ for SiCOH) and introducing porosity (porous SiCOH or pSiCOH, $\kappa = 2.4$)[1,12]. The recommendations of the International Roadmap for Devices and Systems (IRDS) for 2028 call for the urgent development of ultralow-κ dielectrics with κ values of less than 2 (refs. [13,14]). IRDS has also indicated that the greatest challenge concerning interconnect development is the introduction of new materials with reduced dielectric permittivity. Boron-based compounds such as BCN and amorphous

[1]Department of Chemistry, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea. [2]Inorganic Material Lab., Samsung Advanced Institute of Technology (SAIT), Suwon, South Korea. [3]School of Materials Science and Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea. [4]Center for Multidimensional Carbon Materials, Institute for Basic Science (IBS), Ulsan, South Korea. [5]Department of Energy Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea. [6]Pohang Accelerator Laboratory, Gyeongbuk, South Korea. [7]UNIST Central Research Facilities, Ulsan, South Korea. [8]School of Materials Science and Engineering, University of Ulsan, Ulsan, South Korea. [9]Analytical Engineering Group, Samsung Advanced Institute of Technology (SAIT), Suwon, South Korea. [10]Catalan Institute of Nanoscience and Nanotechnology (ICN2), CSIC and BIST, Barcelona, Spain. [11]Institucio Catalana de Recerca i Estudis Avancats (ICREA), Barcelona, Spain. [12]Department of Materials Science & Metallurgy, University of Cambridge, Cambridge, UK. [13]Low-Dimensional Carbon Materials Center, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea. ✉e-mail: mc209@cam.ac.uk; hyeonjin.shin@samsung.com; shin@unist.ac.kr

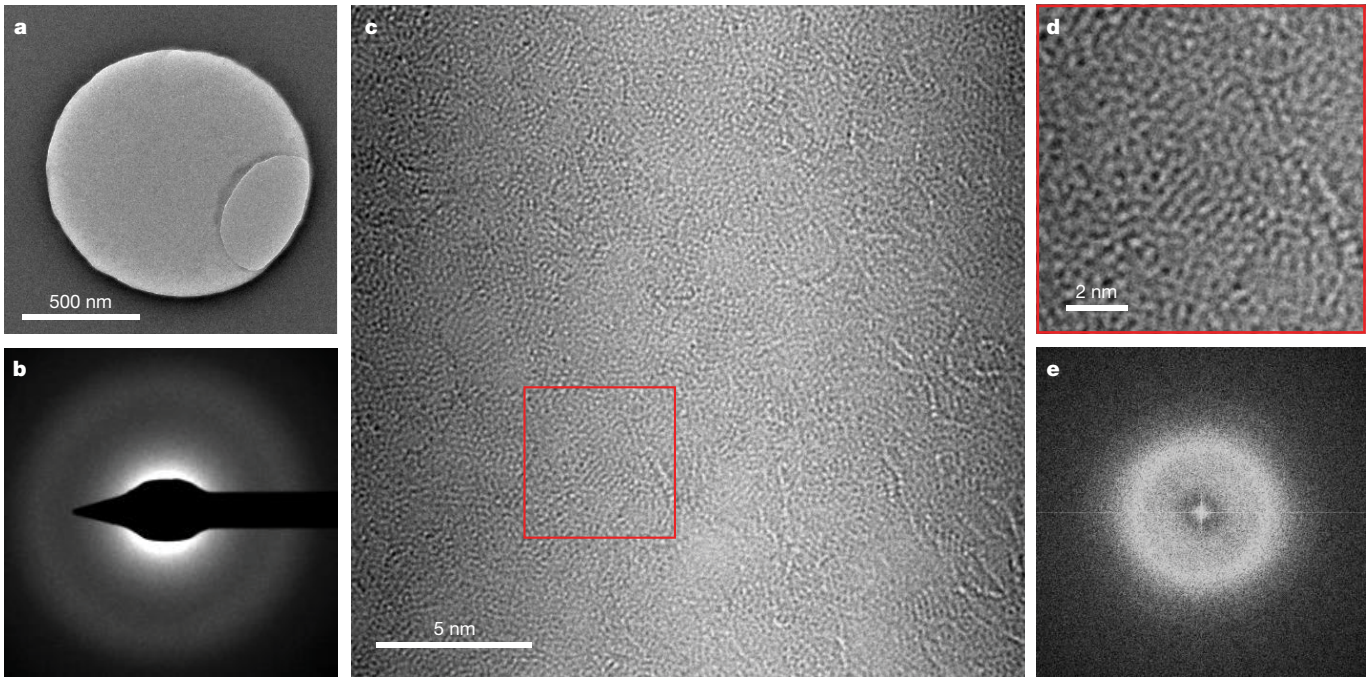


Fig. 1 | Atomic structure of a-BN. a, Low-magnification TEM image of a-BN. **b**, Selected-area electron diffraction image showing a diffuse pattern with no discernible crystalline rings. **c**, High-resolution TEM image. **d**, Magnification of the area indicated by the red box in **c**, showing a disordered atomic arrangement. **e**, Fast Fourier transform results for the area depicted in **d**, demonstrating a diffuse diffraction pattern that is typical of an amorphous film.

boron nitride (a-BN) have been investigated as potential low-κ dielectrics, showing promising results[5]. However, recent reports on BN dielectrics show films with turbostratic structure or high dielectric constants that indicate crystalline structure[15,16]. In this study, we utilized low-temperature remote inductively coupled plasma–chemical vapour deposition (ICP-CVD; see Methods) to obtain 3-nm-thick BN layers on Si substrates (see Extended Data Fig. 1 for growth on Cu and SiO_2 substrates). The transmission electron microscopy imaging and

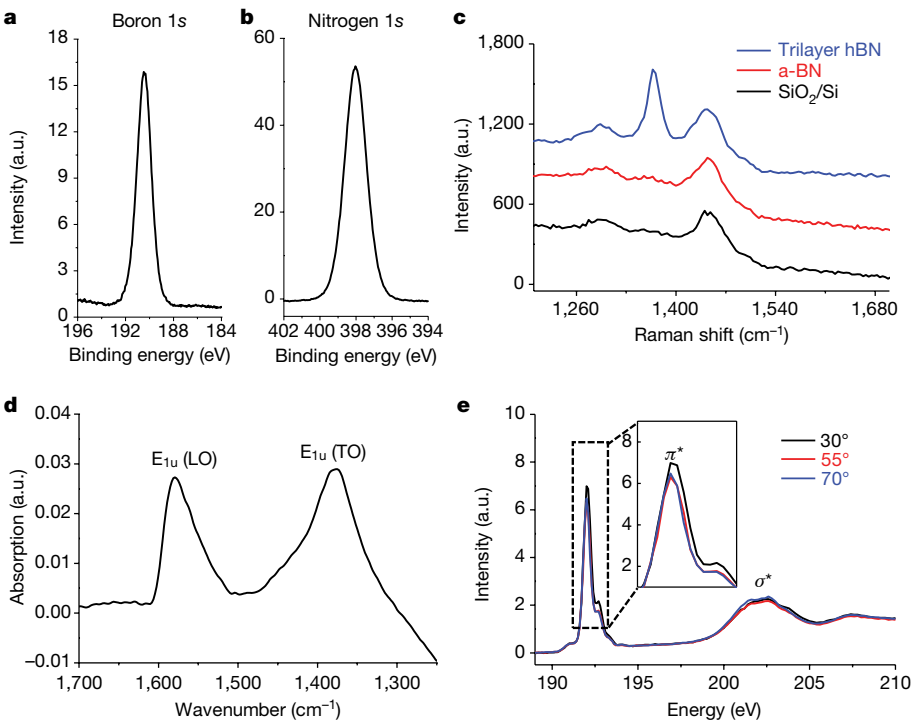


Fig. 2 | Chemical structure of a-BN. a, **b**, XPS profiles for the B 1*s* (**a**) and N 1*s* (**b**) peaks. **c**, Raman spectra of a-BN and epitaxially grown trilayer hBN (used as a reference) on a SiO_2(300 nm)/Si substrate. The Raman spectrum of the SiO_2 substrate is identical to that of a-BN, suggesting that no distinct crystalline hBN modes are present in a-BN. **d**, FTIR spectrum measured using s-polarized radiation at an incident angle of 60°. The absorption peaks near 1,370 cm^{-1} and 1,570 cm^{-1} are attributed to transverse optical (TO) and longitudinal optical (LO) modes of BN, respectively. **e**, PEY-NEXAFS spectra for the B K edge of a-BN, measured at incident angles of 30°, 55° and 70°, showing no dependence on orientation. a.u., arbitrary units.

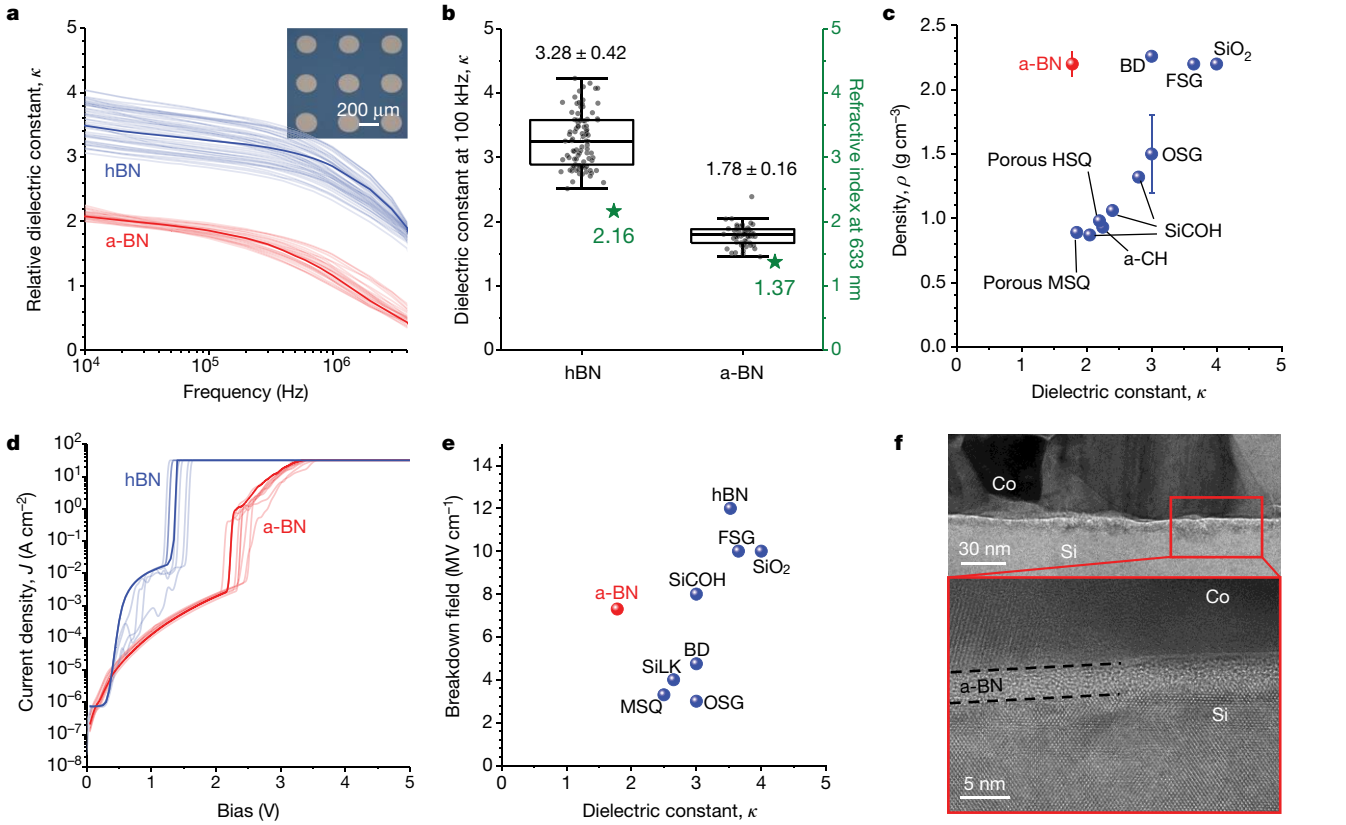


Fig. 3 | Dielectric properties of a-BN. a, Relative dielectric constant as a function of frequency for a-BN and hBN. The dielectric constants were determined by capacitance–frequency measurements on metal–insulator–metal structures (inset). Thick blue and red lines denote averages. **b**, Statistical distribution of dielectric constants measured at 100 kHz and refractive indices (green stars) obtained by ellipsometry for a-BN and hBN. The box indicates a region with a 25% and 75% distribution relative to the average value, and the top and bottom bars mean maximum and minimum values. The number of devices measured in **a** and **b** is 250 for a-BN and 330 for hBN. **c**, Density versus dielectric constant for low-κ materials reported in literature (blue circles) and a-BN (red circle). **d**, Typical current density–voltage curves for hBN (approximately 1.2 nm thick; blue curve) and a-BN (3 nm thick; red curve) films. Thick blue and red lines denote averages. The number of devices measured is 100 for a-BN and 11 for hBN. **e**, Breakdown field versus dielectric constant for low-κ materials reported in the literature (blue circles) and for a-BN (red circle). **f**, Cross-sectional TEM images of Co/a-BN/Si interfaces after annealing for 1 h at 600 °C. The bottom image shows a magnified view of the area marked by the red box in the top image. No diffusion of Co into Si through a-BN is observed. HSQ, hydrogen silsesquioxane; MSQ, methylsilsesquioxane; BD, black diamond; FSG, fluorinated silicon glass; OSG, organosilicate glass; a-CH, amorphous hydrocarbon.

diffraction results shown in Fig. 1 reveal that the films are amorphous with no discernible long-range order; hence, we refer to the material as amorphous BN (a-BN). This is also supported by the analysis of the reduced radial distribution function obtained from the electron diffraction data (see Extended Data Fig. 2a, 2b), which shows broad peaks and a nearest-neighbour distance of 0.144 nm. Cross-sectional chemical mapping confirms that the films consist of B and N (Extended Data Fig. 2c). X-ray photoelectron spectroscopy (XPS) was used to obtain chemical information. The B/N atomic ratio was found to be about 1:1.08 (Fig. 2a, b) with the B 1*s* and N 1*s* peaks at 190.4 eV and 397.9 eV, respectively, indicating that the films are sp^2-bonded B and N (refs. [17,18]). The molecular dynamics simulations shown in Extended Data Fig. 3 confirm the amorphous structure of the BN films, which is consistent with the result in Fig. 1.

Raman spectra of a-BN and crystalline trilayer hexagonal BN (hBN; for comparison) reveal that the hBN E_{2g} mode at 1,373 cm^{-1} is absent in a-BN (Fig. 2c)[17,18]. The Fourier transform infrared spectroscopy (FTIR) spectrum in Fig. 2d shows an absorption peak near 1,370 cm^{-1} that is attributed to the transverse optical mode of BN in a-BN. Another infrared mode located near 1,570 cm^{-1} confirms the amorphous nature of sp^2-bonded BN (ref. [19]). FTIR does not show any N–H or B–H bonds (Extended Data Fig. 4a). Detailed chemical and density analyses were conducted with Rutherford backscattering spectroscopy (RBS) and elastic recoil detection analysis (ERDA) and the results are shown in Extended Data Fig. 4b–d.

Angle-dependent near-edge X-ray absorption fine structure (NEXAFS) measurements in partial electron yield (PEY) mode were made at the Pohang Light Source-II 4D beamline to investigate the chemical and electronic structure of a-BN. In NEXAFS, X-ray absorption is used to excite core electrons of B and N to unoccupied states– that is, 1*s* electrons are excited to empty π^* and/or σ^* states. In the $1s \rightarrow \pi^*$ transition, the spatial orientation of π orbitals strongly affects the transition probability. Thus, information pertaining to the relative orientation of orbitals in hBN layers can be obtained by varying the incidence angle of X-rays[20]. NEXAFS spectra obtained for the a-BN sample at incident angles of 30°, 55° and 70° are shown in Fig. 2e. The observed resonance at 192 eV corresponds to the $1s \rightarrow \pi^*$ transition in boron[20]. The resonance intensity of the $1s \rightarrow \pi^*$ transition in a-BN demonstrates negligible variation with the X-ray incidence angle (Fig. 2e), strongly indicating that BN planes are randomly oriented throughout the material. Similar conclusions can be drawn from NEXAFS spectra of the K edge of N (Extended Data Fig. 4e). Additionally, NEXAFS confirms that a-BN is completely sp^2-hybridized[20,21]. For completeness, we also deposited BN films at different remote ICP-CVD parameters such as power, temperature and pressure. We found that the temperature was the most important parameter, with ideal a-BN film deposition occurring at 400 °C and plasma power of 30 W. Above this temperature we obtained nanocrystalline BN (nc-BN), as shown in Extended Data Fig. 5.

We now discuss the dielectric properties of a-BN. The dielectric constant is a physical measure of how easily electric dipoles can be induced

Table 1 | Properties of a-BN and hBN thin films

	Electrical properties			Film properties	
	Dielectric constant		Breakdown field (MV cm^{-1})	Refractive index, n, at 633 nm	Density (g cm^{-3})
	at 100 kHz	at 1 MHz			
hBN	3.28	2.87	4.0	2.16	2.1
a-BN	1.78	1.16	7.3	1.37	2.1–2.3

in materials by application of an electrical field. The κ value of air or vacuum is 1, but electric polarizability in solid-state matter arises from dipolar, atomic and electronic components that are most relevant for high-performance electronics. The contributions from these can be measured as a function of frequency in the range 10 kHz–30 MHz. The relative dielectric constants (κ) for a-BN and hBN at different frequencies are shown in Fig. 3a. It can be seen that the κ values for hBN and a-BN at 100 kHz are 3.28 and 1.78, respectively. The values are averages of measurements on more than 250 devices. The distribution of the measured values and the corresponding error bars at 100 kHz are shown in Fig. 3b and Table 1. Remarkably, at a frequency of 1 MHz, the observed κ value for a-BN is further reduced to 1.16, which is close to the value of air or vacuum. The low κ values of a-BN are attributed to nonpolar bonds between BN and to the absence of order, which prevents dipole alignment. The κ values for a-BN compare very favourably to other reports in the literature, as shown in Extended Data Table 1. We validated the κ values from the electrical measurements using values obtained by measuring the refractive index (n) of a-BN with spectroscopic ellipsometry and using the relationship $n^2 = \kappa$ (ref. [22]). The refractive indices of hBN and a-BN at a wavelength of 633 nm were found to be 2.16 and 1.37, respectively, as indicated by the green stars in Fig. 3b. Therefore, the κ values for hBN and a-BN from ellipsometry are 4.67 and 1.88, respectively—closely matching those obtained with electrical measurements at 100 kHz. Low-κ dielectric materials are sometimes made porous to exploit the low κ value of air, but this decreases the density of the material, which in turn results in poor mechanical strength. Figure 3c shows that a-BN possesses the lowest dielectric constant at the highest density in comparison with well known low-κ materials (Extended Data Table 1). We also measured the mechanical properties of the a-BN films to confirm their strength. The results of nanoindentation measurements shown in Extended Data Fig. 6a, b indicate that the hardness and stiffness values of the a-BN films are equal to or greater than those of silicon (>11 GPa). The nanoscratch test results shown in Extended Data Fig. 6c also suggest that the films are very well adhered to the substrates.

The electrical breakdown strength of a-BN was extracted by measuring the current density versus the applied bias (Fig. 3d) on vertical sandwich-type devices. The data in Fig. 3d reveal that there is a slight increase in current density due to Poole–Frenkel tunnelling at low voltages, whereas above 2.2 V the leakage current increases sharply, leading to electrical breakdown. Because the thickness of a-BN is 3 nm, the breakdown field is 7.3 MV cm^{-1}—nearly twice that of hBN (see Table 1) and the highest value reported so far for materials with dielectric constants of less than 2, as shown in Fig. 3e. The a-BN film also has an exceptionally low leakage current density of 6.27 μA cm^{-2} at 0.3 V, thus demonstrating its potential for use in 3-nm-node devices. The key dielectric properties of a-BN and hBN are summarized in Table 1.

A key step in the fabrication of back-end-of-line CMOS-compatible logic and memory devices is the deposition of a diffusion barrier between the low-κ dielectric material and the metal wire interconnects to prevent metal atom migration into the insulator. Ideally, this step can be eliminated if the low-κ dielectric material can also serve as the diffusion barrier. We therefore tested the diffusion barrier properties of a-BN by depositing an 80-nm-thick cobalt film on a-BN and annealing the Co/a-BN/Si devices in vacuum for 1 h at 600 °C. These annealing conditions are extremely harsh, and under similar conditions severe diffusion of Co into Si occurs when TiN is used as the barrier layer (Extended Data Fig. 7). By contrast, neither diffusion of Co nor silicide formation was observed in the cross-sectional transmission electron microscopy (TEM) results shown in Fig. 3f (additional data in the form of energy-dispersive spectroscopy composition maps are shown in Extended Data Fig. 8), suggesting that a-BN can serve as both the low-κ dielectric and the diffusion barrier. The comparison of the breakdown bias in the Co/a-BN/Si and Co/TiN/Si devices at various temperatures suggests that the films are stable at high temperatures (Extended Data Fig. 9). Our results indicate that a-BN is an excellent low-κ material for high-performance CMOS electronics.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2375-9.

1. Shamiryan, D., Abell, T., Iacopi, F. & Maex, K. Low-k dielectric materials. *Mater. Today* **7**, 34–39 (2004).
2. Moore's deviation. *Nat. Nanotechnol.* **12**, 1105 (2017).
3. Akinwande, D. et al. Graphene and two-dimensional materials for silicon technology. *Nature* **573**, 507–518 (2019).
4. IRTS. More Moore roadmap in *The International Technology Roadmap for Semiconductors 2.0 White Paper: 2015* 15–16 (IEEE, 2015); http://itrs2.net.
5. King, S. W. Dielectric barrier, etch stop, and metal capping materials for state of the art and beyond metal interconnects. *ECS J. Solid State Sci. Technol.* **4**, N3047 (2015).
6. del Alamo, J. A. Nanometre-scale electronics with III–V compound semiconductors. *Nature* **479**, 317–323 (2011).
7. Venema, L. Silicon electronics and beyond. *Nature* **479**, 309 (2011).
8. Franklin, A. D. Nanomaterials in transistors: from high-performance to thin-film applications. *Science* **349**, aab2750 (2015).
9. Koenderink, A. F., Alù, A. & Polman, A. Nanophotonics: shrinking light-based technology. *Science* **348**, 516–521 (2015).
10. Liu, C. et al. Small footprint transistor architecture for photoswitching logic and in situ memory. *Nat. Nanotechnol.* **14**, 662–667 (2019).
11. Xiang, D., Liu, T. & Chen, W. Fused computing and storage in a 2D transistor. *Nat. Nanotechnol.* **14**, 642–643 (2019).
12. Grill, A. PECVD low and ultralow dielectric constant materials: from invention and research to products. *J. Vac. Sci. Technol. B* **34**, 020801 (2016).
13. IRDS. More Moore in *The International Roadmap for Devices and Systems: 2017*, (IEEE, 2017); http://irds.ieee.org.
14. IRTS. Interconnect in *The International Technology Roadmap for Semiconductors 2.0: 2015* (2015); http://itrs2.net.
15. Liu, J. et al. Plasma deposition of low dielectric constant (k = 2.2-2.4) boron nitride on methylsilsesquioxane-based nanoporous films. *J. Appl. Phys.* **96**, 6679–6684 (2004).
16. Glavin, N. R. et al. Amorphous boron nitride: a universal, ultrathin dielectric for 2D nanoelectronics. *Adv. Funct. Mater.* **26**, 2640–2647 (2016).
17. Kim, G. et al. Growth of high-crystalline, single-layer hexagonal boron nitride on recyclable platinum foil. *Nano Lett.* **13**, 1834–1839 (2013).
18. Jang, A. R. et al. Wafer-scale and wrinkle-free epitaxial growth of single-orientated multilayer hexagonal boron nitride on sapphire. *Nano Lett.* **16**, 3360–3366 (2016).
19. Plass, M. F., Fukarek, W., Mändl, S. & Möller, W. Phase identification of boron nitride thin films by polarized infrared reflection spectroscopy. *Appl. Phys. Lett.* **69**, 46–48 (1996).
20. Kim, D. Y. et al. Role of hydrogen carrier gas on the growth of few layer hexagonal boron nitrides by metal-organic chemical vapor deposition. *AIP Adv.* **7**, 045116 (2017).
21. Jiménez, I. et al. Core-level photoabsorption study of defects and metastable bonding configurations in boron nitride. *Phys. Rev. B* **55**, 12025–12037 (1997).
22. Dharma-wardana, M. W. C. Relation of the refractive index to the dielectric constant containing Doppler-like spatial dispersion. *J. Phys. Math. Gen.* **9**, L93–L97 (1976).

Methods

Cleaning of silicon substrates

The Si substrates were ultrasonicated in acetone for 10 min and subsequently cleaned using iso-propyl alcohol and dried with N_2 gas. Upon completion of the 5-min organic-solvent cleaning process, the Si substrates were subjected to O_2 plasma treatment to remove any carbon impurities remaining on the surface and to make the surface hydrophilic and enhance its wettability. Subsequently, the substrates were immersed in 10% dilute HF solution for 10 min to remove the native oxide. Finally, anhydrous ethanol was used to remove the residual HF solution and the surface was dried using N_2 gas.

Growth of a-BN

The clean Si substrates were placed inside the remote ICP-CVD system (Extended Data Fig. 1) at the centre of a furnace. A borazine (purchased from Gelest) precursor flask was placed in a water bath at −15 °C. The bath temperature before the deposition of a-BN was ramped up to 25 °C. For uniform growth, the substrate was tilted by ~30° using a Cu support. Before ramping up the furnace temperature, the pressure inside the CVD system was reduced to its base value of 1×10^{-4} torr, and 20 standard cubic centimetres (sccm) of H_2 gas was introduced. Subsequently, the furnace temperature was increased at the rate of 10 °C min^{-1} to a set target value (400 °C for a-BN), which was maintained for 20 min before starting the deposition. During growth, plasma generation was performed at a power of 30 W by activating the ICP unit under a flow of borazine gas at 0.05 sccm (controlled by a mass flow controller). Growth was conducted for 90 min. At the end of the deposition, the borazine flow and plasma generation were terminated, and the furnace was cooled to room temperature using 20 sccm of H_2 gas.

Transfer of a-BN films

The a-BN films were transferred to arbitrary substrates using the hydrofluoric acid transfer technique described in ref. [18].

Characterization

Scanning electron microscopy (Verios 460, FEI) and atomic force microscopy (Dimension Icon, Bruker) were used to reveal the surface morphology of the films, and XPS (K-Alpha, Thermo Fisher) was performed to determine their chemical compositions. Raman spectra were measured using a micro Raman spectrometer (alpha 300, WITec GmBH) equipped with a 532-nm laser. To obtain the Raman spectra, samples were transferred onto SiO_2(300 nm)/Si substrates to amplify the signal by multiple reflection[23]. FTIR spectra were acquired using a Varian FTIR 670 spectrometer equipped with a Seagull variable-angle reflection accessory. The incident light was polarized using a wire-grid polarizer. NEXAFS was performed using the 4D PES beamline at the Pohang accelerator laboratory. During NEXAFS, the samples were attached to a molybdenum holder and loaded into a vacuum chamber. The analysis chamber, maintained at a base pressure of 5×10^{-10} torr, was equipped with an electron analyser (R3000, Scienta) and an X-ray absorption spectroscopy detector with a retarding filter to facilitate operation in the PEY mode. For high-resolution imaging and selected-area electron diffraction measurements, low-voltage Cs aberration-corrected TEM (Titan Cube G2 60-300, FEI) was performed at 80 kV using a monochromatic electron beam. To facilitate sample observation using TEM, the a-BN was transferred onto SiN TEM grids (hole diameter, 1 μm). High-resolution cross-sectional TEM (JEM-2100F; JEOL) was performed to confirm the barrier performance.

Molecular dynamics simulations and computations

We modelled the Si substrate using a six-layer diamond rectangular slab having its free surface perpendicular to the *z* axis for BN nucleation and growth. The slab was periodic in the *x*–*y* plane with 18 × 18 repetitions of the unit cell, containing a total of 15,552 Si atoms. The top five layers were completely unrestrained during the simulations, whereas the bottom layer was fixed. The system contained 38,000 atoms of boron and nitrogen at a 1:1 ratio, with an additional 1,900 H atoms (~5%) for consistency with the experimental observations. All the simulations were performed using LAMMPS[24]. Throughout the simulation, the temperature of the substrate was held constant using a Nosé–Hoover thermostat in a canonical *NVT* ensemble at temperature $T = 673$ K. The film was grown using the following method: all the atoms (boron, nitrogen and hydrogen) were initialized with random velocities in a region of height 40 Å above the substrate. They were constantly thermalized at the growth temperature, and allowed to settle and cool on the Si substrate. To prevent premature B–N bond formation, the minimal distance between the initial B and N sources was set at 1.90 Å, larger than the B–N bond length of 1.44 Å in the hBN lattice. The equation of motion was numerically solved using the velocity Verlet integration scheme. Each simulation was run for more than 15 ns at a time step of 0.25 fs. After the growth process, the systems were further relaxed in an NPT ensemble at $T = 300$ K. The extended Tersoff potential for BN was employed to describe the chemical processes (such as bond formation and dissociation) among the atomic species involved[25]. This model potential has been specifically designed to correctly describe the dependence of the bonding in B, N and B–N systems on coordination and chemical environment. Thanks to its versatility, it allows the realization of large-scale atomistic simulations with more than a few thousand atoms. To describe the interaction within the silicon substrate, we used the Tersoff model potential, which has been proved to faithfully reproduce both the mechanical and the morphological properties of silicon-based systems[26]. We treated the Si–N and Si–B interactions using the parameterized[27] Tersoff potential, which has been previously employed to study the compositional and structural features of Si–B–N networks[28]. Finally, we modelled all the interactions involving hydrogen using a Lennard–Jones potential.

Ellipsometry

An automated angle M-2000F rotating-compensator ellipsometer equipped with an *X*–*Y* mapping stage, focusing probes and accompanying software (Complete-EASE 6.39 from J. A. Woollam Co.) was used in this study. Ellipsometric data were acquired in the wavelength range 250–1,000 nm with a resolution of 1.6 nm at incidence angles of 65°, 70° and 75°. The optical properties of both films were determined using the Kramers–Kronig consistent dispersion model using three Lorentz oscillators.

High-resolution Rutherford backscattering and elastic recoil detection analysis

To investigate the elemental composition of the thin films, high-resolution Rutherford backscattering spectrometry (HR-RBS)[29] was performed by irradiating samples with a 450-keV He^+ beam generated by an RBS system (HRBS-V500; Kobe Steel). A magnetic-sector analyser with a high resolution of 1.2 keV was used for the measurements of the thin films. By employing the same system, high-resolution elastic recoil detection analysis (HR-ERDA) was simultaneously performed for hydrogen using 500-keV N^+ ions. Typical beam currents used in the HR-RBS and HR-ERDA analyses were 40 nA and 6 nA, respectively.

Density measurements

Peaks corresponding to the relevant elements (B, N, O and Si) were observed in the HR-RBS spectra. The areas covered by the peaks reflect both the thickness and the density of these elements. The areal density (atoms per centimetre square) was measured[30], enabling the calculation of the a-BN film density by considering the element thickness. Oxygen from surface contamination was observed.

Breakdown voltage and dielectric constant measurements

The current density–voltage (J–V) and capacitance–frequency (C–f) characteristics of the films in metal/a-BN/n-Si stacks were measured using a Tektronics K4200A-SCS parameter analyser system and a Karl Suss PA-200DS semi-automatic probe station. a-BN-based capacitors were fabricated on BN films directly deposited or transferred onto n-Si substrates. To prevent polymer contamination during device fabrication, a shadow mask with a 200-µm-diameter pattern was used, and a 100-nm-thick Cu electrode was deposited over the a-BN/Si stack. After the device fabrication, capacitance–voltage units in the parameter analyser system were used to perform the C–f measurements. We carried out the C–f measurements in the frequency range 1 kHz–10 MHz with a hold bias of 0.5 V and an a.c. drive of ±30 mV. The measured capacitance values did not change substantially as a function of the applied voltage of 0.5 V. Therefore, the relative dielectric constant was evaluated using the relation $\kappa = Ct/A\varepsilon_0$, where t denotes the a-BN film thickness, A represents the area and ε_0 denotes the dielectric constant of vacuum. At high frequencies exceeding 5 MHz, considerable noise levels were observed in the capacitance, probably owing to the low impedance of the a-BN capacitor. Subsequently, the J–V characteristics of both film samples were determined using source measurement units of the parameter analyser system. The applied voltage was swept from 0 to 10 V with a resolution of 1 pA and a compliance current of 10 mA. Additionally, measurements were carried out at 50-mV voltage steps over 10 power line cycles to prevent degradation due to bias stresses.

Diffusion barrier performance

To evaluate the performance of the films as diffusion barriers, ~3-nm-thick samples of a-BN and TiN (deposited by radiofrequency sputtering) were deposited on Si substrates. Subsequently, the samples were coated with 80-nm-thick Co layers using d.c. sputtering. After deposition, the samples were placed inside a furnace for annealing. The furnace temperature was ramped up at a rate of 40 °C min^{-1} in a vacuum of less than 10^{-4} torr. During annealing, thermally activated diffusion is expected to occur at the interface between Co and the dielectric barrier materials.

Data availability

The datasets generated and/or analysed during the current study are available from the corresponding authors on reasonable request.

Code availability

The code used to generate the figures is available from the corresponding authors on reasonable request.

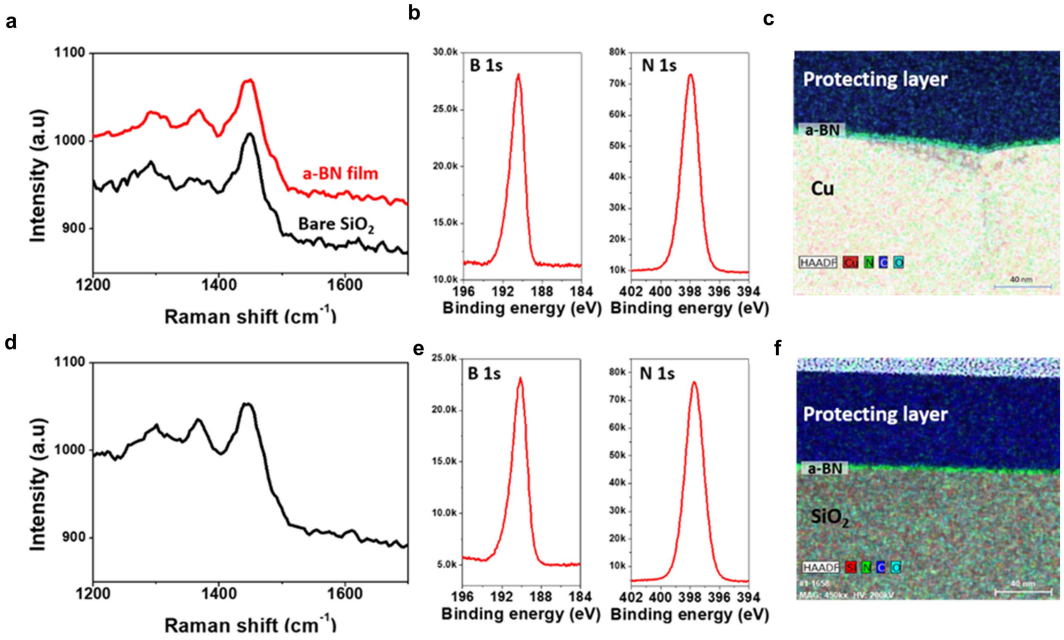
23. Ling, X. & Zhang, J. Interference phenomenon in graphene-enhanced Raman scattering. *J. Phys. Chem. C* **115**, 2835–2840 (2011).
24. Plimpton, S., Crozier, P. & Thompson, A. *LAMMPS-Large-Scale Atomic/Molecular Massively Parallel Simulator* (Sandia National Laboratories, 2007).
25. Los, J. et al. Extended Tersoff potential for boron nitride: energetics and elastic properties of pristine and defective h-BN. *Phys. Rev. B* **96**, 184108 (2017).
26. Tersoff, J. Empirical interatomic potential for silicon with improved elastic properties. *Phys. Rev. B* **38**, 9902 (1988).
27. Matsunaga, K. & Iwamoto, Y. Molecular dynamics study of atomic structure and diffusion behavior in amorphous silicon nitride containing boron. *J. Am. Ceram. Soc.* **84**, 2213–2219 (2001).
28. Al-Ghalith, J., Dasmahapatra, A., Kroll, P., Meletis, E. & Dumitrică, T. Compositional and structural atomistic study of amorphous Si–B–N networks of interest for high-performance coatings. *J. Phys. Chem. C* **120**, 24346–24353 (2016).
29. Kimura, K., Joumori, S., Oota, Y., Nakajima, K. & Suzuki, M. High-resolution RBS: a powerful tool for atomic level characterization. *Nucl. Instrum. Methods Phys. Res. B* **219-220**, 351–357 (2004).
30. Kitahara, A., Yasuno, S. & Fujikawa, K. Study of thin-film thickness and density by high-resolution Rutherford backscattering spectrometry and X-ray reflectivity. *Trans. Mater. Res. Soc. Jpn.* **34**, 613–615 (2009).
31. Hatton, B. D. et al. Materials chemistry for low-*k* materials. *Mater. Today* **9**, 22–31 (2006).
32. Cheng, Y.-L. & Lee, C.-Y. in *Nanoporous Materials Porous Low-Dielectric-Constant Material for Semiconductor Microelectronics* Ch. 6 (IntechOpen, 2018).
33. Zhao, L. et al. Role of copper in time dependent dielectric breakdown of porous organo-silicate glass low-*k* materials. *Appl. Phys. Lett.* **99**, 222110 (2011).
34. Volinsky, A. A., Palacio, M. L. B. & Gerberich, W. W. "Incompressible" pore effect on the mechanical behavior of low-*K* dielectric films. In *MRS Proceedings* Vol. 750, 567–572 (2002).
35. Das, A. et al. Characterisation and integration feasibility of JSR's low-*k* dielectric LKD-5109. *Microelectron. Eng.* **64**, 25–33 (2002).
36. Chang, S. Y. et al. Mechanical property analyses of porous low-dielectric-constant films for stability evaluation of multilevel-interconnect structures. *Thin Solid Films* **460**, 167–174 (2004).
37. Li, H. Y. et al. Process improvement of 0.13 µm Cu/Low K (Black Diamond™) dual damascene interconnection. *Microelectron. Reliab.* **45**, 1134–1143 (2005).
38. Pang. B., Yau W., Lee P. & Naik M. *A New CVD Process for Damascene Low-*k *Applications* 285–289 (Semiconductor Fabtech, 1999).
39. Sekhar, N. V. in *Nanoindentation in Materials Science* (ed. Němeček, J.) Ch. 10 (InTech, 2012).
40. Tyberg, C. et al. in *Polymers for Microelectronics and Nanoelectronics* Vol. 874, 161–172 (American Chemical Society, 2004).
41. Im, J., Townsend, P. H., Curphy, J., Karas, C. & Shaffer, E. O. in *Metallization of Polymers 2* (ed Sacher, E.) 53–60 (Springer, 2002).
42. Grill, A. Plasma enhanced chemical vapor deposited SiCOH dielectrics: from low-*k* to extreme low-*k* interconnect materials. *J. Appl. Phys.* **93**, 1785–1790 (2003).
43. Brandrup, J., Immergut, E. H. & Grulke, E. A. *Polymer Handbook* 4th edn (Wiley-Interscience, 2004).
44. Lee, H. J. et al. Structural characterization of porous low-*k* thin films prepared by different techniques using X-ray porosimetry. *J. Appl. Phys.* **95**, 2355–2359 (2004).
45. Laturia, A., Van de Put, M. L. & Vandenberghe, W. G. Dielectric properties of hexagonal boron nitride and transition metal dichalcogenides: from monolayer to bulk. *npj 2D Mater. Appl.* **2**, 6 (2018).

Acknowledgements We thank UNIST Central Research Facilities (UCRF) and Y. K. Kim for the cross-sectional high-resolution TEM images. This work was supported by Samsung Electronics (Samsung-SKKU Graphene/2D Center), the research fund (NRF-2017R1E1A1A01074493 and NRF-2019R1A4A1027934), the IBS (IBS-R-019-D1) and a grant (CASE-2013M3A6A5073173) from the Centre for Advanced Soft Electronics under the Global Frontier Research Program via the National Research Foundation of the Ministry of Science and ICT, South Korea. The NEXAFS experiments performed at the 4D, 6D and 10A2 beamlines of the Pohang Accelerator Laboratory (PAL) were supported in part by the Ministry of Science and ICT, POSTECH and UNIST. M.C. acknowledges support from Leverhulme Trust Research Grant RPG-2019-227. S.R., A.A. and M.C. acknowledge the European Union Horizon 2020 research and innovation programme for grant number 785219 and 881603 (Graphene Flagship). A.A. is supported by Project MECHANIC (PCI2018-093120) funded by Ministerio de Ciencia, Innovación y Universidades. The Catalan Institute of Nanoscience and Nanotechnology is funded by the CERCA Programme/Generalitat de Catalunya and supported by the Severo Ochoa Centres of Excellence programme, funded by the Spanish Research Agency (grant number SEV-2017-0706).

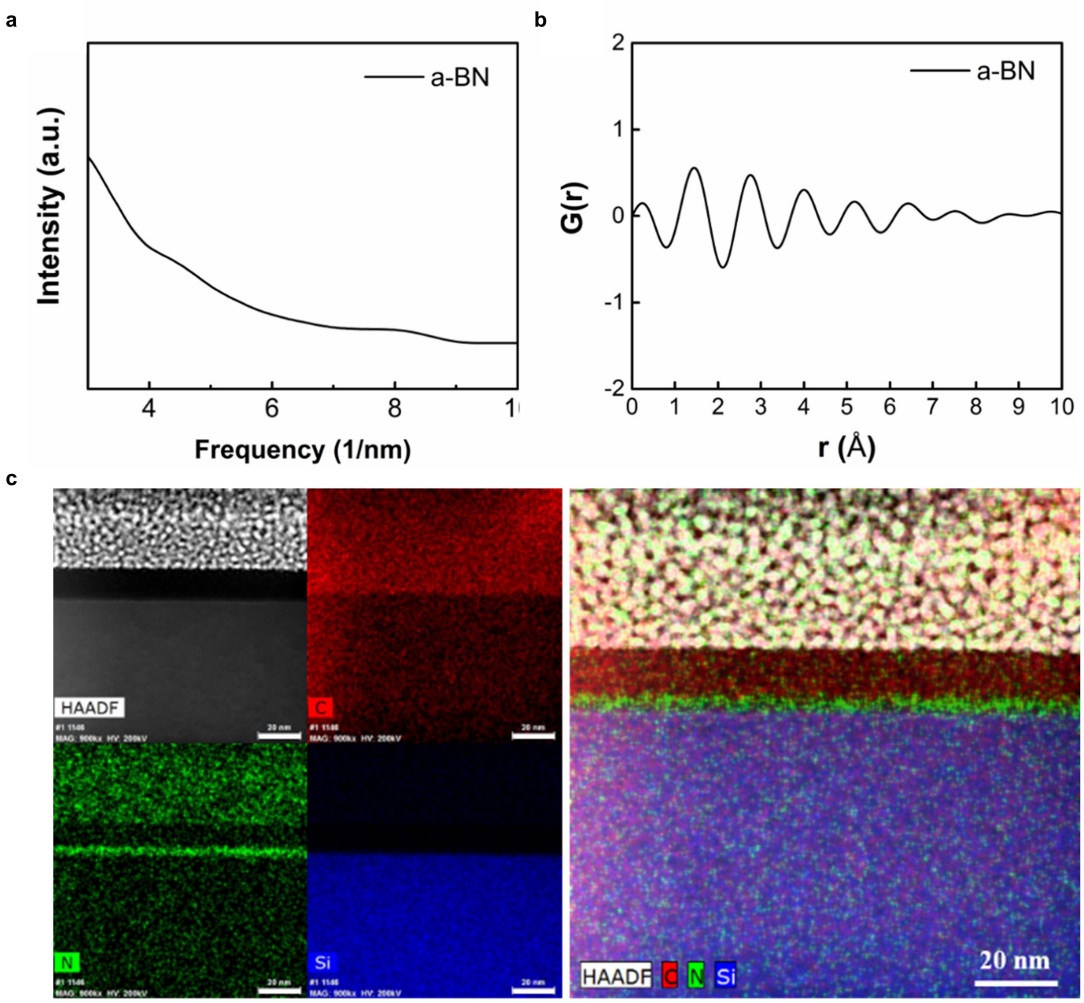
Author contributions H.S.S. and H.-J.S. planned and supervised this project. S.H., K.Y.M., G.K, S.I.Y. and H.S.S. performed the growth and characterization experiments. C.-S.L., M.-H.L. and H.-J.S. fabricated the electrical devices. S.W.K. performed and analysed the ellipsometry measurements. H.-I.L. obtained and analysed the RBS data. Y.L. and Z.L. obtained the TEM data. K.I., K.-J.K. and T.J.S. measured the NEXAFS data. E.-c.J., H.J. and J.-Y.K measured the mechanical properties and adhesion. A.A. and S.R. performed the molecular dynamics simulations. M.C. suggested key experiments and measurements and helped with the interpretation of results. M.C. wrote and edited the manuscript with H.S.S. All authors contributed to the writing of the manuscript and agreed on the contents of the paper.

Competing interests The authors declare no competing interests.

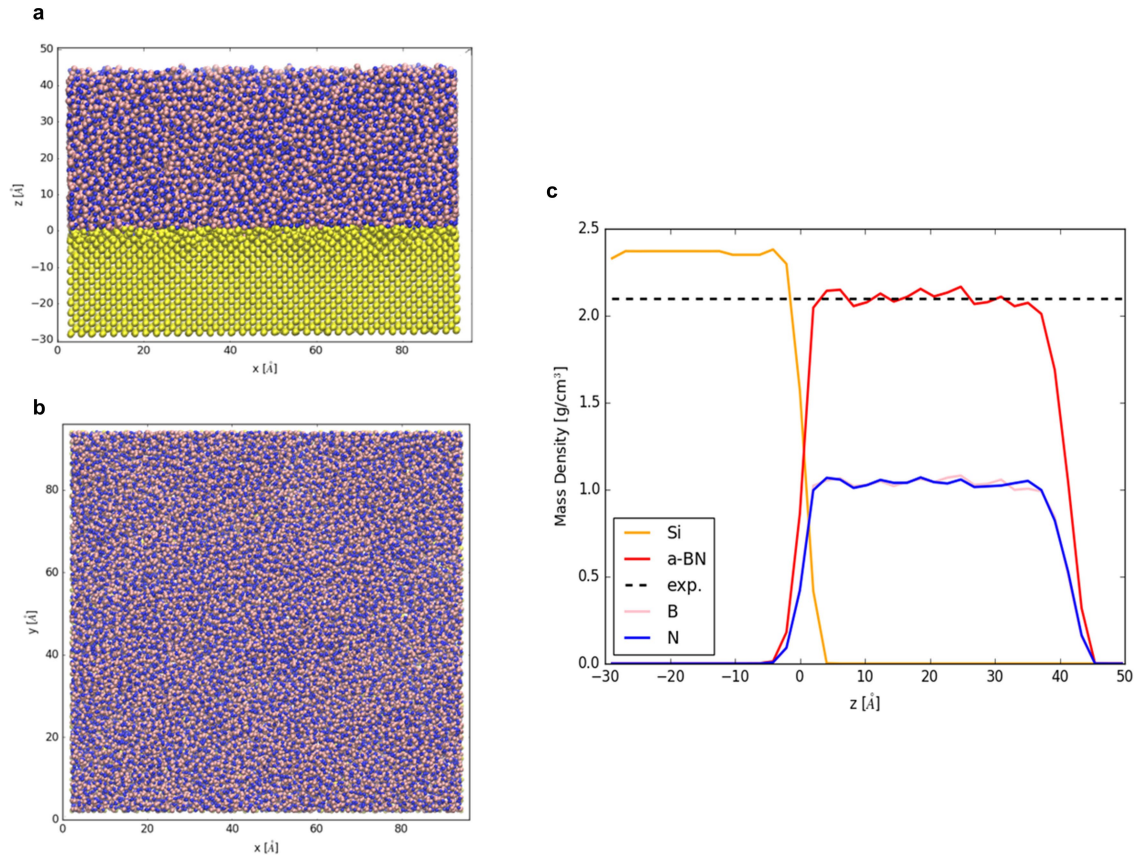
Additional information
Correspondence and requests for materials should be addressed to M.C., H.-J.S. or H.S.S.
Peer review information *Nature* thanks Francesca Iacopi, Lain-Jong Li and Junhao Lin for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



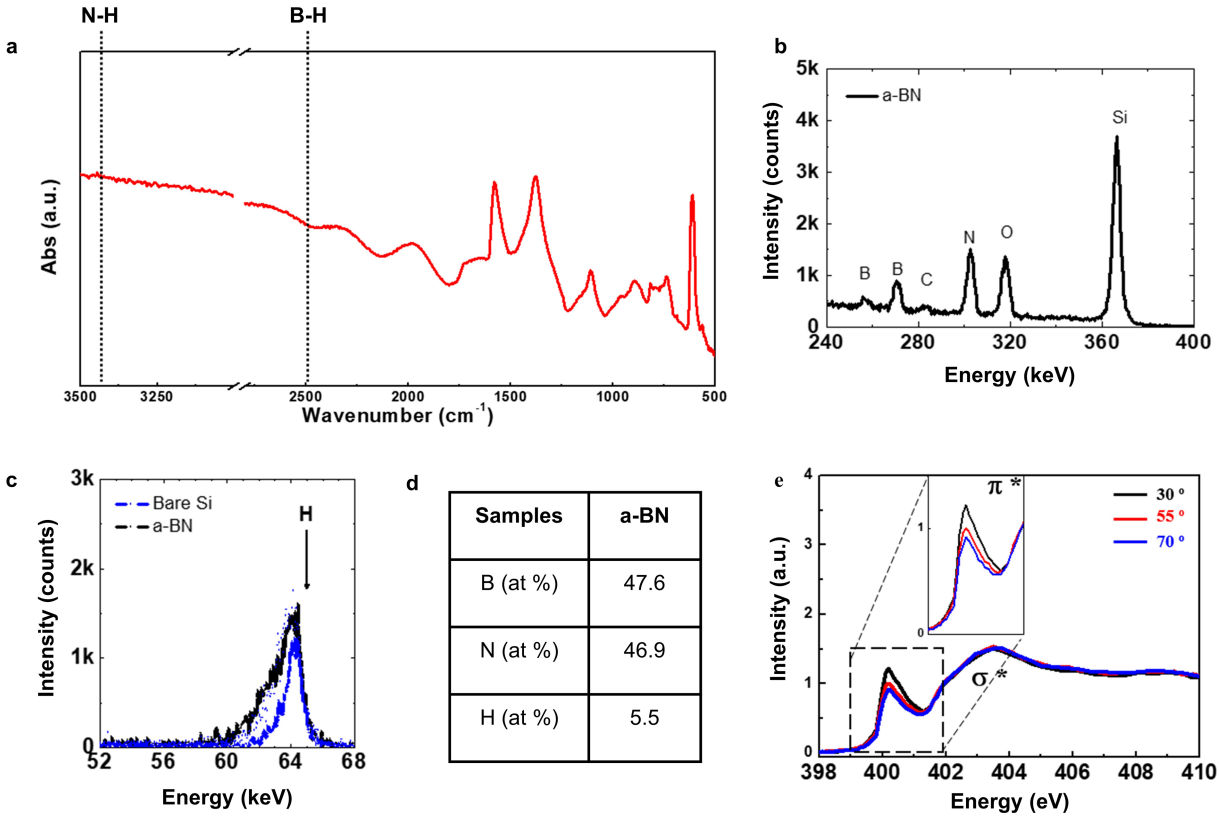
Extended Data Fig. 1 | Growth of a-BN on Cu and SiO_2 substrates. a–**c**, Raman (**a**), XPS (**b**) and EDS mapping (**c**) images of a-BN grown on copper foils (plasma power 30 W, growth temperature 300 °C) and transferred onto SiO_2 substrates for Raman measurements. The typical Raman spectrum of the a-BN film is similar to that of bare amorphous SiO_2. **d**–**f**, Raman (**d**), XPS (**e**) and EDS mapping (**f**) images of an a-BN film grown directly on SiO_2 (plasma power 10 W, growth temperature 200 °C). The spectra are largely the same for all substrates. The dielectric properties obtained from spectroscopic ellipsometry reveal no influence of the substrate. Scale bar, 40 nm.



Extended Data Fig. 2 | Analysis of the reduced radial distribution function obtained from the electron diffraction data and cross-sectional chemical mapping of the a-BN film. a, Azimuthally averaged experimental electron diffraction intensity of a-BN. **b**, Reduced radial distribution function, $G(r)$, of a-BN obtained from the electron diffraction data. The peak position $r = 1.44$ Å corresponds to the nearest-neighbour distance of B–N. The $G(r)$ curve was calculated using eRDF Analyser (an open-source interactive GUI for electron reduced density function analysis). **c**, High-angle annular dark-field (HAADF) scanning TEM image (left) overlaid with EDS maps of carbon (red), nitrogen (green) and silicon (blue). An image with overlaid EDS maps for all elements is shown on the right. Scale bars, 20 nm.

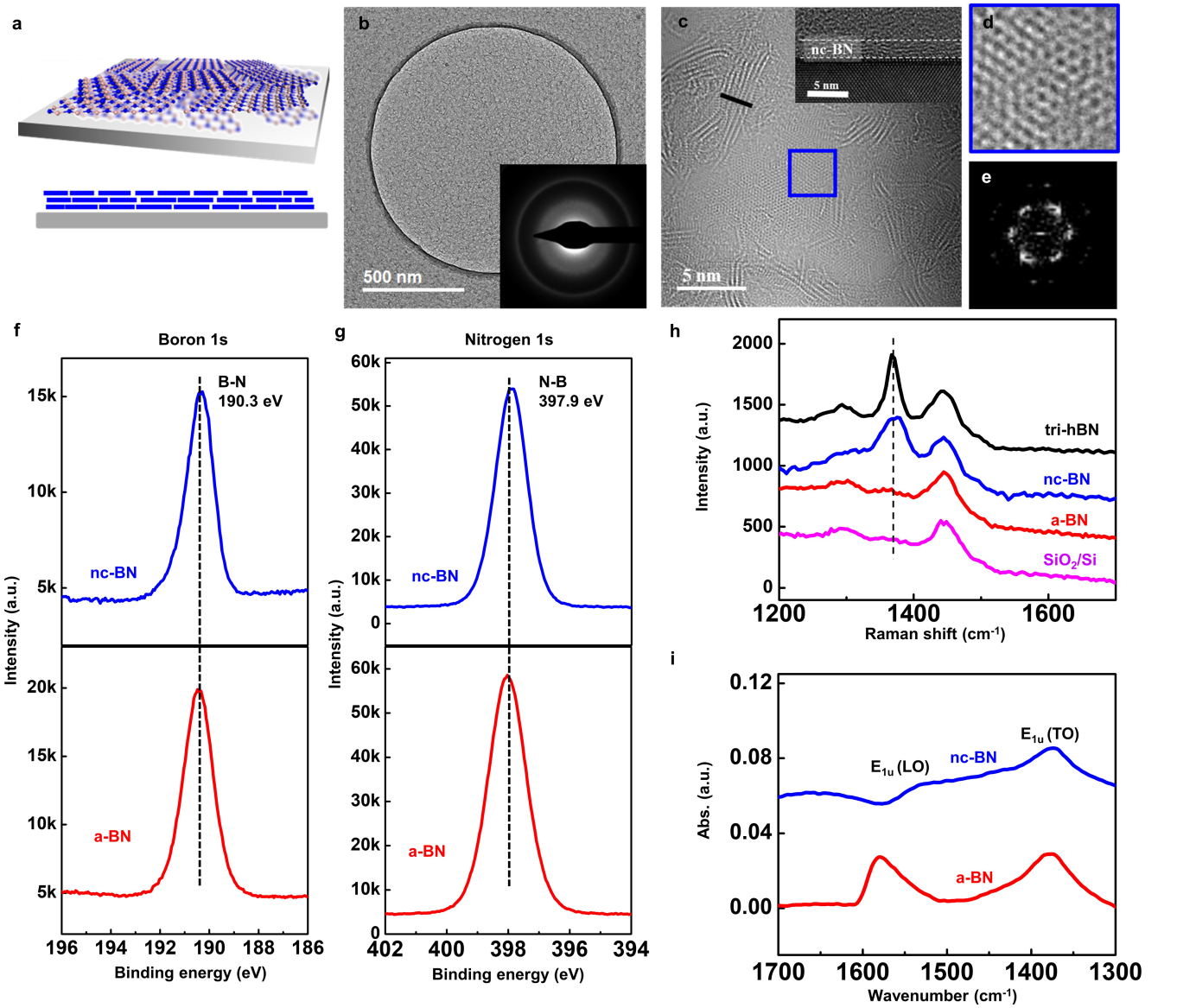


Extended Data Fig. 3 | Molecular dynamics simulation. a, **b**, Side view (**a**) and top view (**b**) of a-BN grown on Si substrates at 673 K, calculated using molecular dynamics simulations. Different atomic species are shown in different colours: yellow (Si), blue (N) and pink (B). **c**, Mass density profile along the transverse direction (*z*), obtained from the results shown in **a** and **b**. Coloured solid lines denote the densities of different chemical species. The simulated density of a-BN is consistent with the experimental result. The black dashed line corresponds to the measured BN mass density.

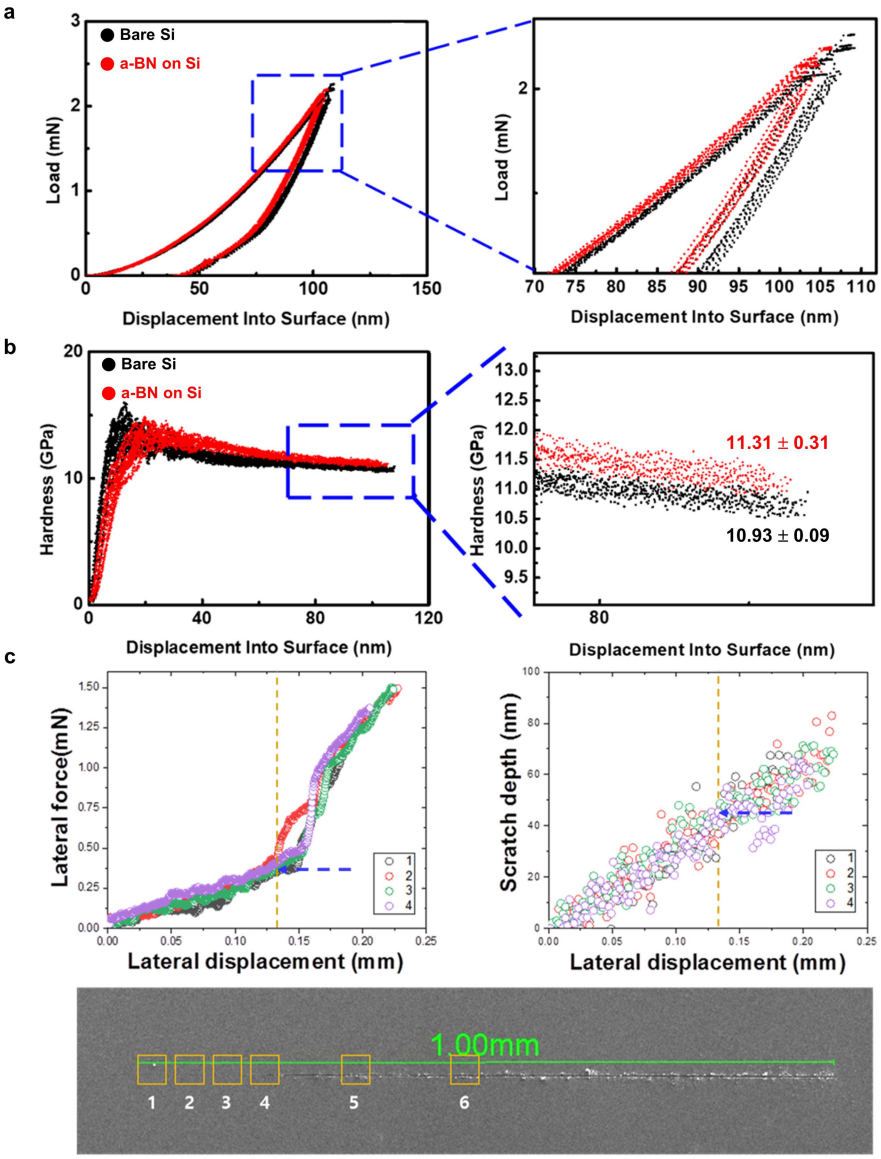


Samples	a-BN
B (at %)	47.6
N (at %)	46.9
H (at %)	5.5

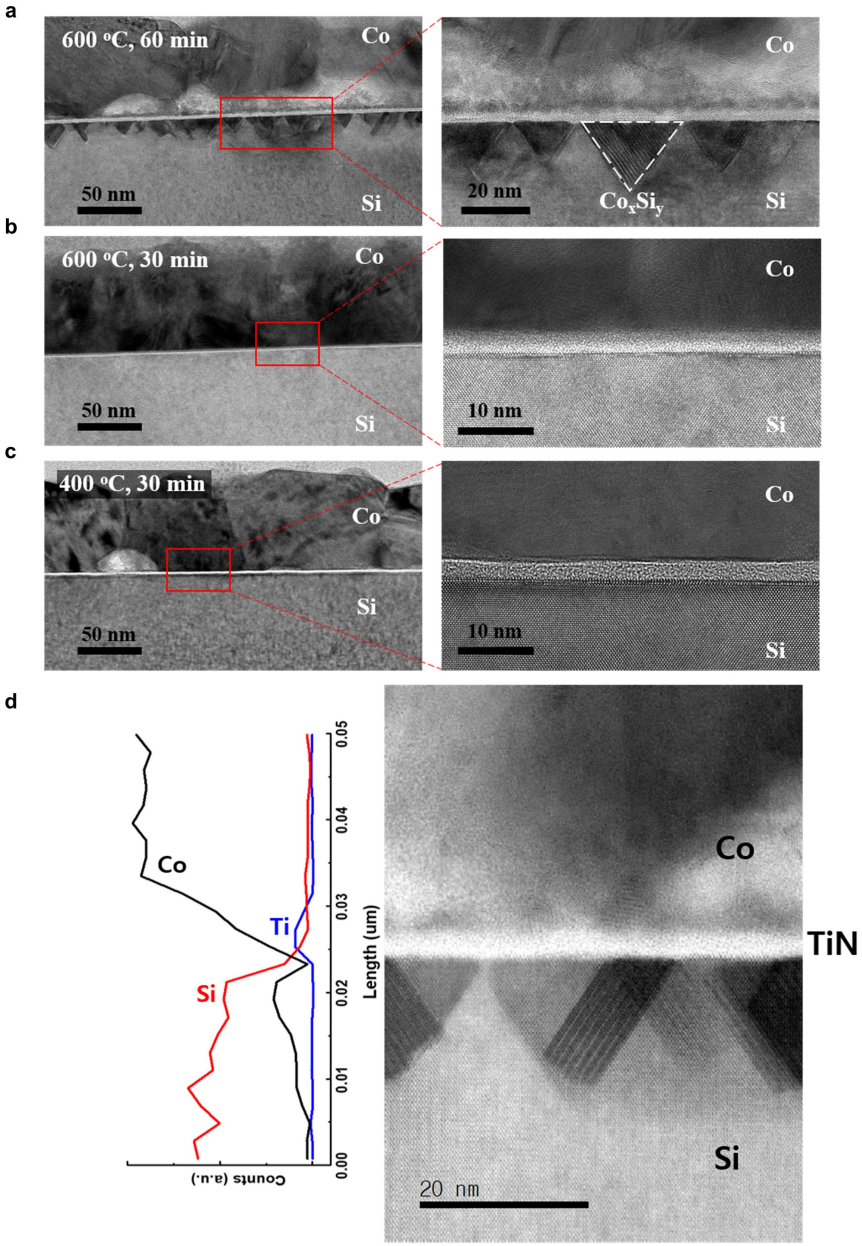
Extended Data Fig. 4 | FTIR, HR-RBS, HR-ERDA and NEXAFS analyses of a-BN films. a, FTIR spectra of a-BN, showing the absence of B–H and N–H bonds. Abs, absorption. **b**, **c**, HR-RBS (**b**) and HR-ERDA (**c**) spectra of an a-BN film in the energy range 240–400 keV and 52–68 keV, respectively. **d**, Elemental composition calculated using the HR-RBS and HR-ERDA spectra. **e**, PEY-NEXAFS spectra for the N K edge of a-BN, measured at incident angles of 30°, 55° and 70°, demonstrating a small angular dependence of the N K edge.



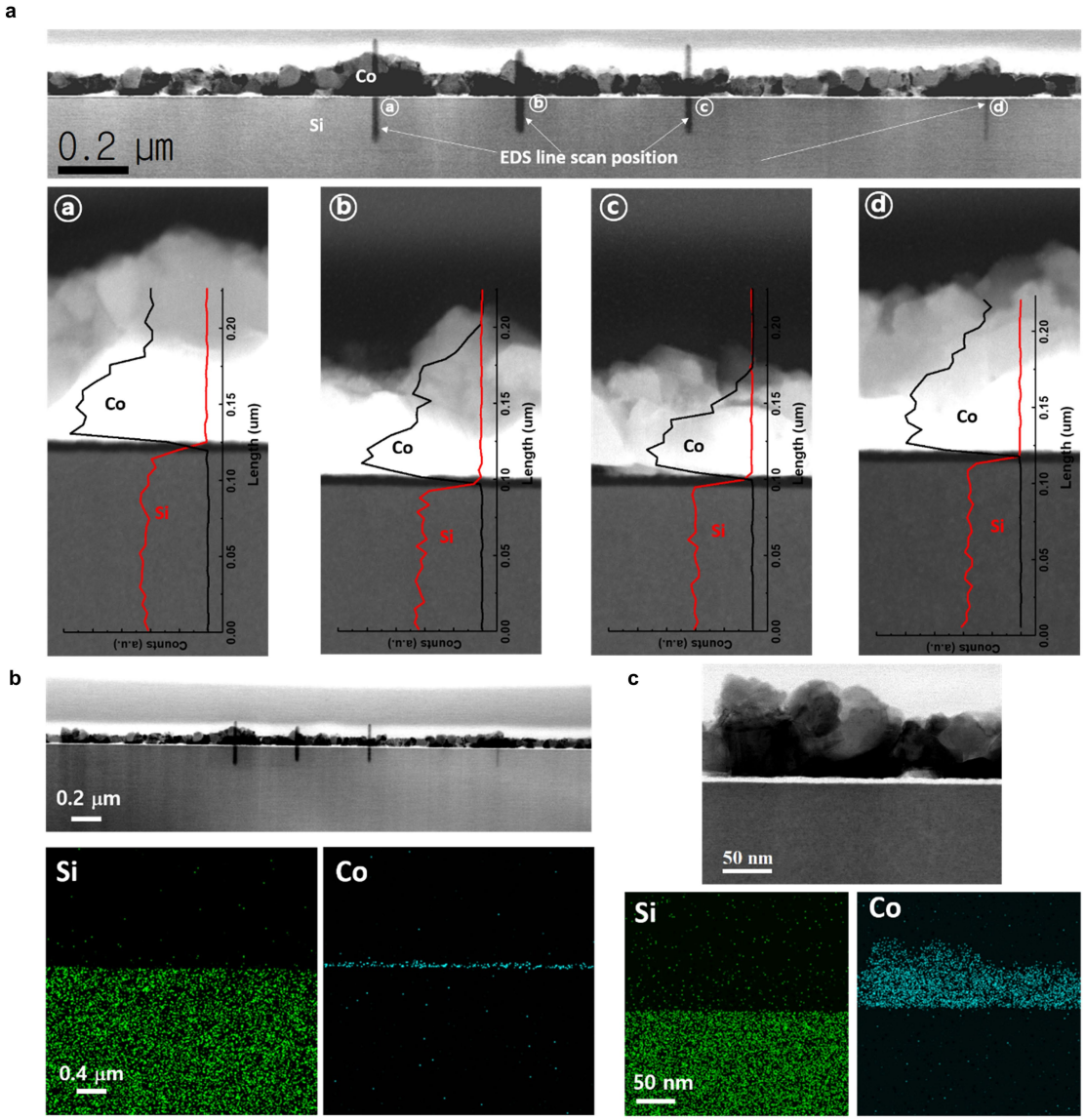
Extended Data Fig. 5 | Comparison of a-BN and nc-BN films. a, Structure of nc-BN film deposited at 700 °C. **b**, Low-magnification TEM images of nc-BN. The selected-area electron diffraction pattern in the inset shows a typical polycrystalline ring pattern. **c**, High-resolution TEM images of nc-BN, clearly showing small crystallites of hBN. The cross-sectional TEM image in the inset indicates a layered structure. **d**, Magnification of the area indicated by the blue box in **c**. **e**, Fast Fourier transform image showing the hexagonal superstructure of multilayer hBN. **f**, **g**, XPS profiles of the B 1*s* (**f**) and N 1*s* (**g**) peaks observed in 3-nm-thick a-BN and nc-BN samples. **h**, Raman spectra of a-BN, nc-BN and epitaxially grown trilayer hBN (tri-hBN; 1.2 nm thick; used as reference) samples transferred onto SiO_2/Si substrates. **i**, FTIR spectra for a-BN (red) and nc-BN (blue) measured using s-polarized radiation at an incident angle of 60°. The E_{1u} longitudinal optical (LO) mode is related to the amorphous phase of BN; see ref. [19].



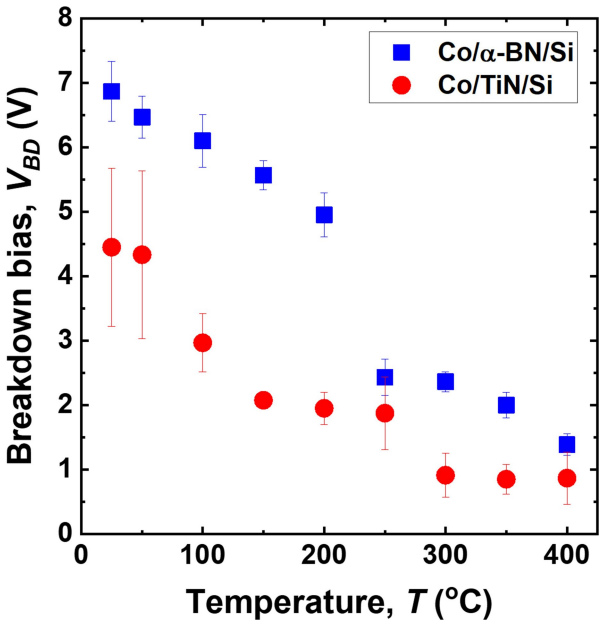
Extended Data Fig. 6 | Nanoindentation and nanoscratch test results. **a**, **b**, Nanoindentation results showing that the deposition of a-BN on Si substrates leads to an enhancement in surface hardness and stiffness. The average and standard deviation values for the hardness in **b** are 11.31 ± 0.13 for a-BN on Si and 10.93 ± 0.09 for bare Si. **c**, Nanoscratch test results revealing that a scratch depth of 40 nm (more than 10 times the film thickness) is required to delaminate the film, which suggests excellent adhesion with the Si substrate. Scanning electron microscopy observations show that the scratch regions are clean, with no evidence of delamination of a-BN for a scratch depth smaller than 40 nm. The datasets 1–4 in the bottom panel represent scratch test data obtained at four different positions under the same experimental conditions.



Extended Data Fig. 7 | Cross-sectional TEM images of Co(80 nm)/TiN(3 nm)/Si films after thermal diffusion tests at different temperatures. a–**c**, Images obtained after thermal diffusion at 600 °C for 60 min (**a**), 600 °C for 30 min (**b**) and 400 °C for 30 min (**c**). **d**, Magnified cross-sectional TEM image (right) and EDS line profile (left) of the film shown in **a**.



Extended Data Fig. 8 | Large-area cross-sectional TEM images, EDS line profiles and maps of a Co/a-BN(3 nm)/Si film after thermal diffusion at 600 °C for 60 min. a, Large-area cross-sectional TEM image and EDS line profiles. **b**, EDS maps of Co and Si showing that Co is isolated above the a-BN film and does not diffuse into the Si. **c**, EDS maps of a magnified area in **b**.



Extended Data Fig. 9 | Breakdown bias at different temperatures for a-BN and TiN barriers.

Extended Data Table 1 | Comparison of dielectric constants of various dielectric materials

	Dielectric constant	Density (g/cm^3)	Modulus (GPa)	Hardness (GPa)	Breakdown field (MV/cm)	
SiO_2	4	2.2	55~70	3.5	>10	Ref. [31,32]
FSG (Fluorinated silicon glass)	3.5~3.8	2.2	~50	3.36	>10	Ref. [31,32]
OSG (organosilicate glass or carbon-doped silicon glass)	2.8~3.2	1.8~1.2	6.6~8.4	1.2~1.7	3.0 Ref. [33]	Ref. [31,32,34]
HSQ	~3.0					
MSQ	~2.5		2.7~12.5	0.19~0.936	3.3 Ref. [35]	Ref. [36]
Black Diamond (SiCOH)	2.7~3.3	<2.26	10~20	1.5~3.0	4.75 Ref. [37]	Ref.[38,39]
$Si_wC_xO_yH_z$	2.7~3.0		9~15	1.3~2.4	6~10	Ref. [40]
SiLK	2.65		2.7	0.38	4	Ref.[40,41]
SiCOH	2.8	1.32	16.2	1.69		Ref. [42]
pSiCOH (pore <1.5 nm)	2.4	1.06	4.2	0.28		Ref. [42]
pSiCOH (pore <2.5 nm)	2.05	0.87	3.3	0.28		Ref. [42]
$a\text{-}CH_{polymeric}$	2.2~2.3	0.92~0.94				Ref.[34,43]
Porous HSQ (hydrogen silsesquioxane) (porosity 46%)	2.2	0.98				Ref. [44]
Porous MSQ (methylsilsesquioxane) (porosity 34%)	1.85	0.89				Ref. [44]
BCN (boron carbon nitride)	3.7~4.6					Ref. [5]
h-BN	3.29~3.76	2.1	19.5~100	0.6~3		Ref. [45]
a-BN or amorphous h-BN	2.2~2.4 5.9					Ref.[15] Ref.[16]
a-BN	1.16~1.78	2.1~2.3			7.3	This work

Literature data from refs. [5,15,16,31–45]. SiLK, low-κ polymer from Dow Chemicals.

Article

High-strength Damascus steel by additive manufacturing

https://doi.org/10.1038/s41586-020-2409-3

Received: 14 November 2019

Accepted: 27 March 2020

Published online: 24 June 2020

Check for updates

Philipp Kürnsteiner[1]✉, Markus Benjamin Wilms[2], Andreas Weisheit[2], Baptiste Gault[1,3], Eric Aimé Jägle[1,4] & Dierk Raabe[1]

Laser additive manufacturing is attractive for the production of complex, three-dimensional parts from metallic powder using a computer-aided design model[1–3]. The approach enables the digital control of the processing parameters and thus the resulting alloy's microstructure, for example, by using high cooling rates and cyclic re-heating[4–10]. We recently showed that this cyclic re-heating, the so-called intrinsic heat treatment, can trigger nickel-aluminium precipitation in an iron–nickel–aluminium alloy in situ during laser additive manufacturing[9]. Here we report a Fe19Ni5Ti (weight per cent) steel tailor-designed for laser additive manufacturing. This steel is hardened in situ by nickel-titanium nanoprecipitation, and martensite is also formed in situ, starting at a readily accessible temperature of 200 degrees Celsius. Local control of both the nanoprecipitation and the martensitic transformation during the fabrication leads to complex microstructure hierarchies across multiple length scales, from approximately 100-micrometre-thick layers down to nanoscale precipitates. Inspired by ancient Damascus steels[11–14]—which have hard and soft layers, originally introduced via the folding and forging techniques of skilled blacksmiths—we produced a material consisting of alternating soft and hard layers. Our material has a tensile strength of 1,300 megapascals and 10 per cent elongation, showing superior mechanical properties to those of ancient Damascus steel[12]. The principles of in situ precipitation strengthening and local microstructure control used here can be applied to a wide range of precipitation-hardened alloys and different additive manufacturing processes.

Parts built by laser additive manufacturing (LAM) experience a specific thermal history. First comes a rapid quenching from the liquid state, followed by an intrinsic heat treatment (IHT), that is, cyclic re-heating that consists of a multitude of short temperature spikes[6,9,15]. In directed energy deposition (DED), parts are built layer-wise by laser melting of powder fed by a carrier gas through a nozzle[1,3]. In DED, the IHT is pronounced and hence provides opportunities to locally adjust the microstructures[7–9,15,16]. Yet new materials must be tailor-designed to best exploit these specific conditions, as conventional alloy compositions cannot be expected to perform efficiently as they have been optimized for other processing routes, for example, casting or forging.

We recently showed that the IHT can trigger nickel-aluminium (NiAl) precipitation in an iron–nickel–aluminium (Fe–Ni–Al) alloy[9]. This so-called maraging steel draws properties from two important phase transformations. Initially, a soft nickel-rich martensitic microstructure forms upon quenching through an austenite-to-martensite transformation. This martensite is later hardened by a second phase transformation to form intermetallic nanoprecipitates. Therefore, conventionally produced as well as LAM-produced commercial maraging steels (for example, 18Ni-300) need to undergo a costly ageing treatment to form property-enhancing intermetallic precipitates[17–22]. Iron–nickel–titanium (Fe–Ni–Ti) alloy systems show extremely fast kinetics for Ni_3Ti precipitation[17–19,21–23], making them ideally suited for in situ precipitation hardening by exploiting the short temperature peaks during IHT.

Here the digital control of the DED process parameters allowed us to locally exploit these two phase transformations and adjust the microstructure to create a new material inspired by Damascus steel. The layered structure of Damascus steel originally resulted from repeatedly folding and forging macrocomposites consisting of a hard steel and a soft steel and lends excellent strength and ductility to the composite[11–14]. We utilize this concept here and produce a Damascus-like maraging steel, not by folding and forging but by fabricating layered microstructures by exploiting rapid quenching, sequential in situ heating and local phase transformation. We specifically designed an Fe19Ni5Ti (wt%) alloy to exploit the rapid quenching and IHT of the DED (see Methods section Supplementary discussion of the alloy design concept' for more details). We adjusted the DED process parameters to regulate the time–temperature profile during the fabrication process, which enabled precise, local control of the martensite formation as well as the

[1]Department Microstructure Physics and Alloy Design, Max-Planck-Institut für Eisenforschung, Düsseldorf, Germany. [2]Fraunhofer Institute for Laser Technology ILT, Aachen, Germany. [3]Department of Materials, Royal School of Mines, Imperial College London, London, UK. [4]Present address: Institute of Materials Science, Universität der Bundeswehr München, Neubiberg, Germany. ✉e-mail: p.kuernsteiner@mpie.de

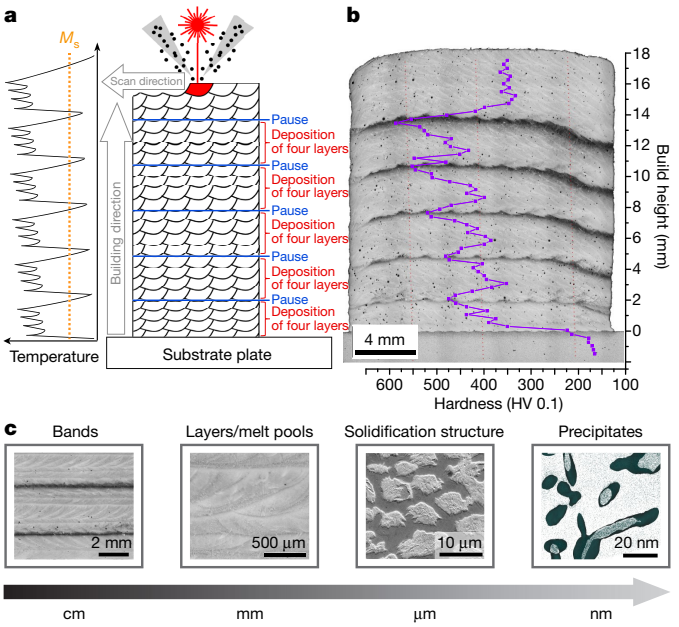


Fig. 1 | DED-produced Fe19Ni5Ti (wt%) sample. a, A schematic of the DED process including a simple sketch of the temperature profile that also shows the martensite start temperature M_s. After building four layers in a sequence, the process was paused for 120 s, allowing the sample to cool. **b**, A light optical micrograph showing a dark band at the position where a pause was introduced. The overlay of the hardness curve shows a peak in hardness at each dark band. **c**, An overview of the hierarchy of the microstructural features at different length scales that are discussed throughout this paper.

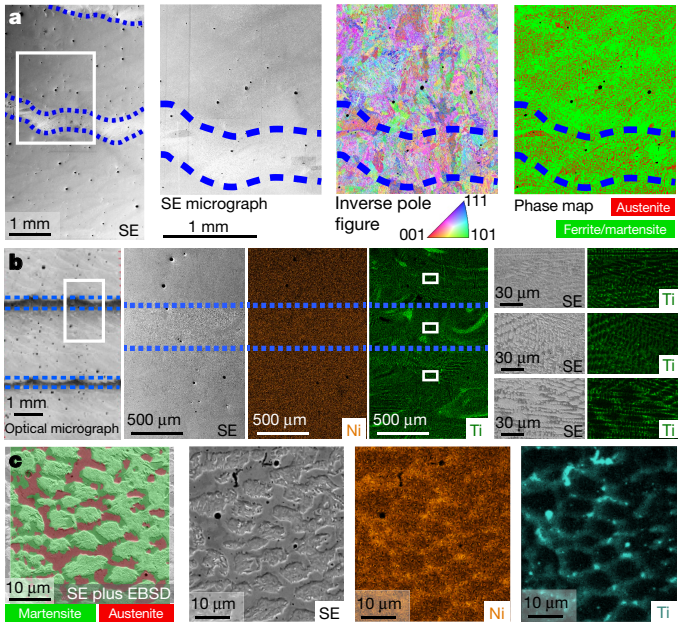


Fig. 2 | Microstructure characterization at different length scales. The hard regions/bands are marked by dashed blue lines. These hard regions appear dark in the optical micrographs because of the rough surface that scatters light away. However, this rougher surface emits a higher number of secondary electrons (SEs), which lets the hard bands appear bright in the electron micrographs. The white boxes mark the area that is magnified in the images to the right. **a**, An SE micrograph featuring two hard regions and the corresponding EBSD maps across one hard band. The inverse pole figure map is in the sample build direction. **b**, Two hard regions in an optical micrograph plus elemental mapping of Ni and Ti at different magnifications. **c**, An SE micrograph with an EBSD map overlaid and the corresponding elemental mapping of Ni and Ti at higher magnification than **b**. Inhomogeneities in the Ti distribution result from the mixing of Ti powder particles due to Marangoni convection as well microsegregation that results in austenite stabilization at interdendritic regions.

precipitation, and hence of the mechanical behaviour. Our approach avoids a time-consuming and costly post-process ageing heat treatment, and also provides the possibility to locally tune the microstructure, which would not be possible with conventional heat treatments.

Results and discussion

Overview

We built a cuboidal part with Fe19Ni5Ti (wt%) maraging steel by DED. DED, as a LAM approach, uses a computer-controlled deposition strategy that, here, included a 120-s pause after a block of four layers. During this pause, the laser was switched off and the sample cooled. A schematic of the process and a micrograph are shown in Fig. 1. The pause led to the formation of a dark band at the top of each block that was deposited continuously without pause. The superimposed hardness profile shows that the dark bands are approximately 100 HV harder than the intermediate four-layer blocks. Both the dark contrast (further illustrated in Extended Data Fig. 1) and the increased hardness result from a precipitation reaction discussed in detail in the next section. These dark bands, on a millimetre–centimetre length scale, represent the coarsest constituent of the hierarchical microstructure of our Damascus-steel sketched in Fig. 1c.

Microstructure analysis

Figure 2 shows a microstructure typical of LAM-produced maraging steels, consisting of a Ni-martensitic matrix with retained austenite occurring in the interdendritic regions. Austenite is stabilized there because the interdendritic regions are enriched in solutes[17,21,22]. Electron backscatter diffraction (EBSD) showed that both the hard bands and the softer regions have a similar austenite fraction and martensite morphology (Fig. 2a). Elemental mapping revealed inhomogeneities in the Ti distribution on two different length scales (Fig. 2b, c). (1) Incomplete mixing of the pre-alloyed Fe20Ni (wt%) powder with elemental Ti powder during fluid flow in the melt pool (Marangoni convection) leads to Ti-enriched regions that are a few hundred micrometres in size. These mixing inhomogeneities do not have an influence on the overall phase fractions. (2) Microsegregation of Ti to the interdendritic regions during solidification leads to micrometre-sized regions enriched in Ti. Figure 2b shows that there are no discernible differences in the distribution or concentration of alloying elements between the hard regions and the soft regions.

Figure 2c illustrates the role of Ti and Ni microsegregation in stabilizing the austenite. The electron micrograph with the overlaid EBSD map shows that the smooth, darker-appearing areas are austenite. Martensite appears brighter because of the rougher surface emitting more secondary electrons. The elemental mapping indicates that the austenite in the interdendritic regions is enriched in Ti and Ni (see Extended Data Fig. 2 for more details of the microstructure characterization at that length scale). This is counterintuitive because Ti is usually classified as a ferrite-stabilizing element in steels. However, we calculated the driving force for martensite formation using Calculation of Phase Diagrams (CALPHAD) simulations, which showed that Ti enrichment lowers the Gibbs energy difference between the austenite and martensite (Extended Data Fig. 3). In this alloy, Ti hence acts as an austenite stabilizer. These dendritic and interdendritic regions resulting from the rapid cooling during DED represent the intermediate constituent of the hierarchical microstructure illustrated in Fig. 1c.

Figure 3a, b shows the fabricated material and a 5-nm-thick slice through an atom probe tomography (APT) reconstruction from analyses of the austenite and martensite in the soft region. Only Ti atoms are shown and they appear randomly distributed in both phases. In

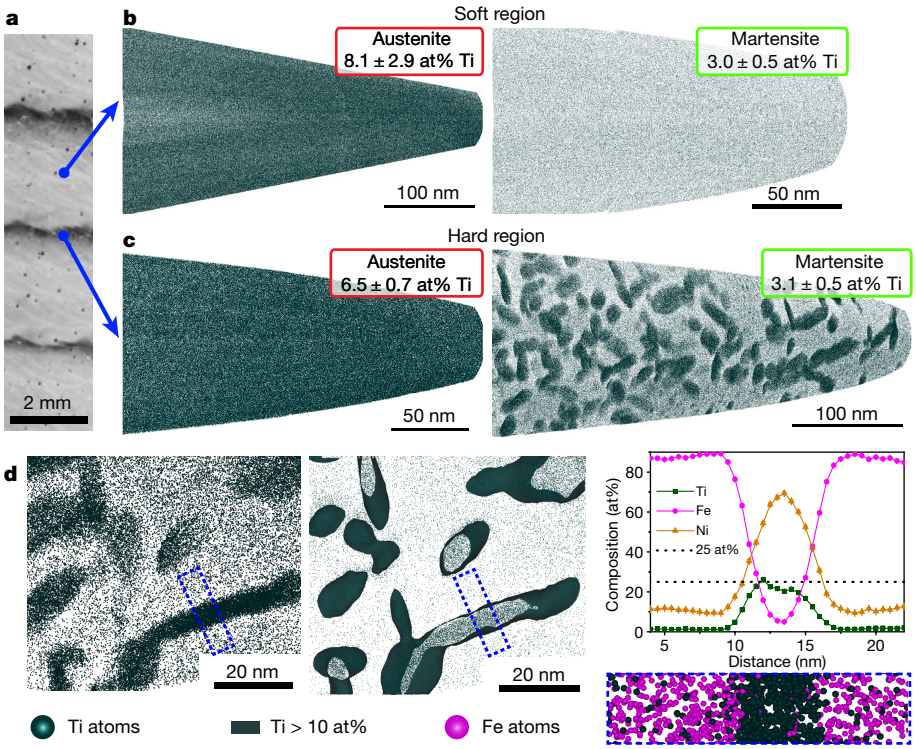


Fig. 3 | APT analysis of martensite and austenite in the soft region and hard region. a, An optical micrograph indicating the positions at which the APT analysis was performed. **b**, **c**, Ti atom maps of a 5-nm-thick slice through the reconstructed volume are shown for the soft region (**b**) and for the hard region (**c**). The left maps shows APT reconstructions from austenite and the right maps show those from martensite. Only the martensite phase forms precipitates upon IHT and only in the hard region. In the soft region, both phases are free of precipitates. **d**, A magnified view of the precipitates by means of Ti atom maps (left) and isocomposition surfaces encompassing regions containing more than 10 at% Ti in dark green (middle). The precipitate is the η-phase $(Ni,Fe)_3Ti$, as can be seen from the one-dimensional composition profile (right) across a precipitate along the dashed blue rectangle. The dashed blue rectangle below the graph shows a magnified view of the rectangles to the left depicting both Fe and Ti atoms as pink and dark green spheres, respectively.

the hard region, however, a high density of precipitates appears in the martensite (Fig. 3c). In this region, two successive phase transformations have occurred: first, an austenite-to-martensite transformation, and second, precipitation inside the martensite. Precipitation occurs only in martensite as the solubility of the alloying elements is higher in austenite. Extended Data Fig. 4 and Supplementary Videos 1 and 2 further illustrate the complex network of precipitates found in this martensite. Averaging over multiple APT datasets, the precipitate volume fraction was determined as 3.50 ± 0.51%.

Precipitates in the Ti atom map in Fig. 3d are highlighted by a set of isocomposition surfaces encompassing regions containing more than 10 at% Ti (dark green). The composition profile across a single plate-shaped precipitate shows that the composition is compatible with η-type Ni_3Ti (for further compositional analysis, see Extended Data Fig. 5). These η-phase precipitates represent the intended precipitate phase selected in the design of our Damascus-like steel. They are the smallest microstructure constituent (Fig. 1c).

Thermal history

Whether η-phase precipitation occurs or not (that is, dark band versus bright region in between) is dictated by the thermal history of the sample during DED: martensite forms only if the temperature drops below the martensite start temperature M_s, and the subsequent cyclic re-heating of the IHT can trigger precipitation (Fig. 4a). The expected phases in this steel appear in the phase diagram isopleth in Extended Data Fig. 6. Figure 4b shows experimental time–temperature profiles from the top surface acquired with a pyrometer during DED using identical process parameters but different pause times after each fourth layer. Without pause time (red solid line), the temperature increases continuously and does not allow the austenite formed upon solidification to transform to martensite. In this case, martensite forms only during the final cool-down after DED, and with no further IHT, no hard bands form (Fig. 4c). In contrast, for samples with a pause in laser illumination, the material cools after each block of four layers. The M_s of the DED-produced Fe19Ni5Ti (wt%) was determined to be 195 °C by dilatometry experiments (Extended Data Fig. 7). During the pause, the temperature drops below M_s and the material of the four layers deposited continuously transforms to martensite (first phase transformation). The subsequent temperature spikes of the IHT trigger η-phase precipitation (second phase transformation). The resulting dark, precipitate-hardened regions appear dark in the optical micrographs in Fig. 4c.

The crucial parameter determining whether precipitation is triggered by the IHT is the temperature drop during the pause (that is, valleys in Fig. 4b). For the sample built with a 30-s pause, the temperature still gradually increases over time, that is, with increasing build height. During the last pause, the temperature only drops to around 180 °C, that is, barely below M_s (orange arrow in Fig. 4b). Only a small fraction of the austenite likely transformed into martensite, and no dark region is discernible for this last pause (that is, the topmost dark region is 'missing', see orange arrow in Fig. 4c). For a sample built with 90-s pauses (shown in Extended Data Fig. 8), the temperature drops below M_s at each pause. However, there is still a slight overall increase in temperature during the build time, causing higher temperature spikes during IHT, which becomes more effective and triggers precipitation to a greater depth into the block of four layers that transformed to martensite during the pause (leading to the broader dark regions towards the top of this sample). For pauses of 180 s (Fig. 4b), the sample cools

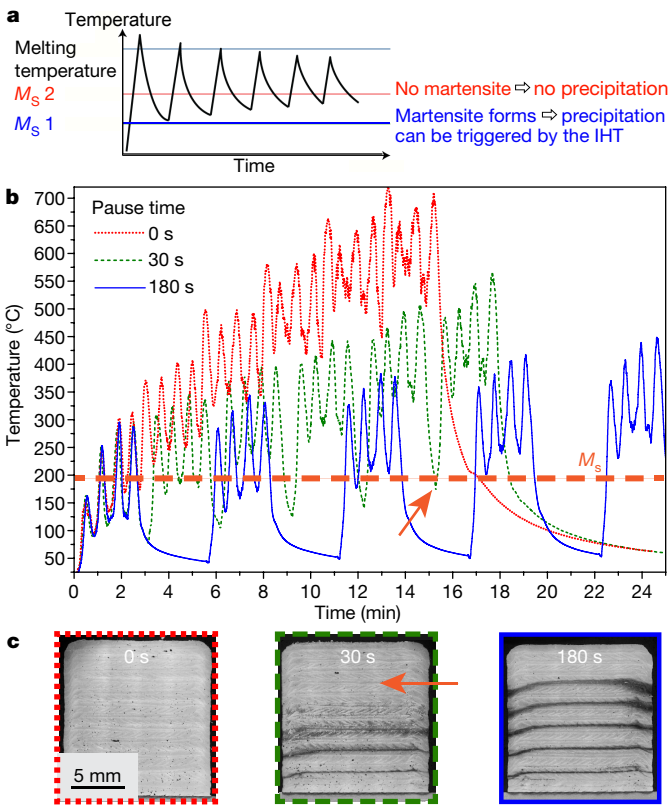


Fig. 4 | The effect of the thermal history. a, A schematic of the IHT. Only if the temperature has dropped below M_s can precipitation be triggered upon the IHT in the martensite phase. **b**, Experimental time–temperature profiles acquired with a pyrometer on the surface of the sample during the DED build at different pause times after each fourth layer. It becomes apparent that without pauses, the temperature increases during the entire fabrication and only drops notably when a pause time is introduced. The dashed orange line corresponds to M_s. The orange arrow points at a temperature drop that barely drops below M_s. **c**, Optical micrographs of the samples built with the corresponding pause times.

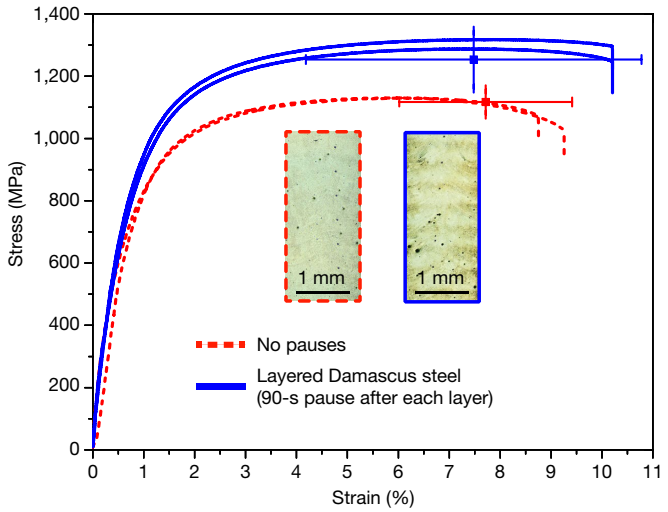


Fig. 5 | Tensile tests of two Fe19Ni5Ti (wt%) steel samples. One Damascus-like sample containing precipitation-hardened bands and a sample containing no precipitates. Only the sample with the pause time (90 s in this case) can cool below M_s during the process and therefore contains martensite hardened by $(Fe,Ni)_3Ti$ precipitates. Two representative curves are shown for each condition together with the average value for maximum tensile strength and elongation at fracture. The insets show the corresponding optical micrographs.

to room temperature during each pause. The spikes in temperature during the IHT have a similar height from the bottom to the top of the sample, which results in narrow, dark, precipitation-hardened bands with a constant thickness of a few hundred micrometres.

Mechanical properties

To probe the tensile properties of the digitally engineered Damascus-like microstructures, we prepared a sample deliberately without any pause, which is devoid of precipitates, and a sample with a pause time of 90 s after each layer. This latter Damascus-type steel has a layered structure featuring dark bands at each DED layer. As a pause was introduced after each layer instead of after each fourth layer as in the previous samples, the pause time was shorter. Tensile curves in Fig. 5 show a substantial increase in yield strength and ultimate tensile strength of approximately 200 MPa. Further results, including outliers, are shown in Extended Data Fig. 9a. Interestingly, the precipitation-hardened, layered sample shows not only an increased strength but also an increased elongation at fracture. The simultaneous increase in strength and ductility presumably stems from the Damascus-like, layered microstructure. Further tensile tests in the build direction, reflecting the mechanical anisotropy of the material, are shown in Extended Data Fig. 9b. Impact toughness can be found in Extended Data Fig. 10.

Considering the simplicity of the ternary Fe19Ni5Ti (wt%) alloy, the ultimate tensile strength of greater than 1,300 MPa paired with >10% elongation compares well to complex conventional 18-Ni300 (1.2709) maraging steels produced by LAM: they reach around 1,000–1,200 MPa ultimate tensile strength with 8–12% elongation in the as-produced state and 1,800–2,100 MPa with 1.5–5.0% elongation in the aged condition[17,24,25]. Conventionally produced 18-Ni300 reaches similar tensile strengths of around 2,000 MPa but has a slightly higher elongation at fracture of around 10%.

Conclusions

We have shown here that hierarchically structured Damascus-like metallic composites can be directly synthesized in situ by additive manufacturing using digital control of the IHT sequences associated with the layer-wise fabrication technique. More specifically, we used a nanostructured martensitic (maraging) steel. Using controlled pausing between alternating layers, we built a composite microstructure with excellent mechanical properties (1,300 MPa and 10% elongation). Its structure consists of mesoscopic soft regions, that is, devoid of nano-precipitates, and hard regions containing a high volume fraction of nanoscale precipitates. These precipitates form over the course of the IHT following the martensitic transformation, which is itself triggered during the cooling offered by the pause. This achievement was enabled by the design of an Fe19Ni5Ti (wt%) alloy specifically for LAM that allows us to tune the start temperature of the martensite transformation and hence the precipitation during the process.

Here we have chosen to vary the pause time between layers because its influence on the temperature is very intuitive and measurable. The local sample temperature can, however, be controlled by a variety of process parameters such as laser power, scan speed, external heating and cooling, and so on, or a combination thereof. This makes the approach presented here applicable to a wide range of additive manufacturing processes. Furthermore, in situ hardening exploiting the IHT can be extended to other precipitation-hardening alloys. The opportunity to locally tailor microstructures and mechanical properties provides new possibilities for manufacturing. As an example, one could manufacture tools that are soft and tough on the inside and only the outer skin is precipitation hardened without the need to apply a coating or a case-hardening treatment.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2409-3.

1. Herzog, D., Seyda, V., Wycisk, E. & Emmelmann, C. Additive manufacturing of metals. *Acta Mater*. **117**, 371–392 (2016).
2. Sames, W. J., List, F. A., Pannala, S., Dehoff, R. R. & Babu, S. S. The metallurgy and processing science of metal additive manufacturing. *Int. Mater. Rev*. **61**, 315–360 (2016).
3. Gu, D. D., Meiners, W., Wissenbach, K. & Poprawe, R. Laser additive manufacturing of metallic components: materials, processes and mechanisms. *Int. Mater. Rev*. **57**, 133–164 (2012).
4. Xu, W., Lui, E. W., Pateras, A., Qian, M. & Brandt, M. In situ tailoring microstructure in additively manufactured Ti-6Al-4V for superior mechanical performance. *Acta Mater*. **125**, 390–400 (2017).
5. Haubrich, J. et al. The role of lattice defects, element partitioning and intrinsic heat effects on the microstructure in selective laser melted Ti-6Al-4V. *Acta Mater*. **167**, 136–148 (2019).
6. Krakhmalev, P., Yadroitsava, I., Fredriksson, G. & Yadroitsev, I. In situ heat treatment in selective laser melted martensitic AISI 420 stainless steels. *Mater. Des*. **87**, 380–385 (2015).
7. Mazumder, J., Choi, J., Nagarathnam, K., Koch, J. & Hetzner, D. The direct metal deposition of H13 tool steel for 3-D components. *JOM* **49**, 55–60 (1997).
8. Cottam, R., Wang, J. & Luzin, V. Characterization of microstructure and residual stress in a 3D H13 tool steel component produced by additive manufacturing. *J. Mater. Res*. **29**, 1978–1986 (2014).
9. Kürnsteiner, P. et al. Massive nanoprecipitation in an Fe-19Ni-*x*Al maraging steel triggered by the intrinsic heat treatment during laser metal deposition. *Acta Mater*. **129**, 52–60 (2017).
10. Costa, L., Vilar, R., Reti, T. & Deus, A. M. Rapid tooling by laser powder deposition: process simulation using finite element analysis. *Acta Mater*. **53**, 3987–3999 (2005).
11. Mintách, R., Nový, F., Bokůvka, O. & Chalupová, M. Impact strength and failure analysis of welded Damascus steel. *Mater. Eng*. **19**, 22–28 (2012).
12. Peterson, D. T., Baker, H. H. & Verhoeven, J. D. Damascus steel, characterization of one Damascus steel sword. *Mater. Charact*. **24**, 355–374 (1990).
13. Verhoeven, J. D. Genuine Damascus steel: a type of banded microstructure in hypereutectoid steels. *Steel Res*. **73**, 356–365 (2002).
14. Černý, M., Filípek, J., Mazal, P. & Dostál, P. Basic mechanical properties of layered steels. *Acta Univ. Agric. Silvic. Mendel. Brun*. **61**, 25–38 (2013).
15. Zheng, B., Zhou, Y., Smugeresky, J. E., Schoenung, J. M. & Lavernia, E. J. Thermal behavior and microstructural evolution during laser deposition with laser-engineered net shaping: Part I. Numerical calculations. *Metall. Mater. Trans. A* **39**, 2228–2236 (2008).
16. Kürnsteiner, P. et al. Control of thermally stable core–shell nano-precipitates in additively manufactured Al-Sc-Zr alloys. *Addit. Manuf*. **32**, 100910 (2020).
17. Bajaj, P. et al. Steels in additive manufacturing: a review of their microstructure and properties. *Mater. Sci. Eng. A* **722**, 138633 (2020).
18. Sha, W., Cerezo, A. & Smith, G. D. W. Phase chemistry and precipitation reactions in maraging steels: Part IV. Discussion and conclusions. *Metall. Trans. A* **24**, 1251–1256 (1993).
19. Pereloma, E. V., Shekhter, A., Miller, M. K. & Ringer, S. P. Ageing behaviour of an Fe-20Ni-1.8Mn-1.6Ti-0.59Al (wt%) maraging alloy: clustering, precipitation and hardening. *Acta Mater*. **52**, 5589–5602 (2004).
20. Bodziak, S. et al. Precipitation in 300 grade maraging steel built by selective laser melting: aging at 510 °C for 2 h. *Mater. Charact*. **151**, 73–83 (2019).
21. Jägle, E. A. et al. Comparison of maraging steel micro- and nanostructure produced conventionally and by laser additive manufacturing. *Materials* **10**, 8 (2017).
22. Tan, C. et al. Microstructural evolution, nanoprecipitation behavior and mechanical properties of selective laser melted high-performance grade 300 maraging steel. *Mater. Des*. **134**, 23–34 (2017).
23. Pereloma, E. V., Stohr, R. A., Miller, M. K. & Ringer, S. P. Observation of precipitation evolution in Fe-Ni-Mn-Ti-Al maraging steel by atom probe tomography. *Metall. Mater. Trans*. **40**, 3069–3075 (2009).
24. Casati, R., Lemke, J., Tuissi, A. & Vedani, M. Aging behaviour and mechanical performance of 18-Ni 300 steel processed by selective laser melting. *Metals* **6**, 218 (2016).
25. Hermann Becker, T. & Dimitrov, D. The achievable mechanical properties of SLM produced maraging Steel 300 components. *Rapid Prototyp. J*. **22**, 487–494 (2016).

Methods

Laser additive manufacturing

Pre-alloyed Fe20Ni (wt%) powder mixed with commercially pure Ti powder to obtain Fe19Ni5Ti (wt%) was used to manufacture samples for this study by DED. Both powders were gas atomized under argon. The Fe20Ni powder was purchased from Nanoval and the pure Ti powder was purchased from TLS Technik. Both powders had a particle fraction of 45–90 µm.

For DED, we used a five-axis handling system equipped with a fibre-coupled diode laser system LDM 3000-60 (Laserline) with a wavelength of 976 nm and a beam parameter product of 60 mm mrad. The final beam diameter of 1.8 mm was obtained through a collimation lens (focal length f_c = 65 mm) and a focusing lens (f_f = 195 mm). A disk-based powder feeding system Sulzer Metco Twin 10C (OC Oerlikon) was used to feed the dry mixed Fe20Ni powder and Ti powder. Argon was used as both the shielding gas and the carrier gas. We applied a bidirectional scan strategy, depositing 20 single tracks with a constant track offset of 0.8 mm (in-plane) with a deposition speed of 600 mm min^{-1} and a laser power of 550 W. The optical setup was moved after completion of each layer by a constant height offset of 0.62 mm (plane to plane) in the building direction. Samples were built on 1.2365 (AISI H10) steel substrate plates. While the aforementioned parameters were kept constant among all samples in this study, we exploited the flexibility of the computer-controlled process to vary a single process parameter, namely, the interlayer pause time. The main sample of this study on which we carried out an in-depth microstructure analysis was produced with 120-s pause time after each fourth layer. This means that after depositing a 'block' of four layers with continuous laser illumination, the process was interrupted for 120 s, during which no heat was imposed by the laser. To measure the cooling during this pause time, samples with different pause times between 0 and 180 s after each fourth layer were produced and the temperature was monitored in situ with a pyrometer. The layered, Damascus-type steel sample with a pause time of 90 s after each layer was selected for in-depth mechanical property characterization, by both tensile and impact testing. The mechanical properties were compared with a sample built without any pause. The pause time of 90 s was chosen to assure cooling of the material below M_s after each layer was deposited, to trigger in situ precipitation with the IHT exerted by the subsequent build layer. Compared with the sample with a pause only after each fourth layer, this sample required a shorter pause time to achieve sufficient cooling because of the more frequent pauses. It is noted that the interlayer pause time is a typical LAM process parameter that is set digitally together with all the other parameters and input into the DED machine.

A LaserSight (Optris) infrared pyrometer was used to acquire the time–temperature profiles during the DED build. After each layer, the pyrometer was moved upwards the same distance as the DED layer height.

Analytical methods

Scanning electron microscopy (SEM) to obtain electron micrographs as well energy dispersive X-ray spectroscopy (EDS) to obtain element mappings was performed in a Zeiss Merlin (Carl Zeiss SMT) featuring a Gemini 2-type field emission gun (FEG) electron column. For EDS, a Brucker XFlash 6/30 silicon drift detector featuring a 30-mm^2 detector area was used. EBSD was performed on a Zeiss 1540XB cross-beam SEM-focused ion beam (FIB) setup featuring a Gemini 1 FEG electron column. EBSD was performed using an EDAX Hikari camera. The TSL OIM Analysis software (version 7) was used for EBSD data analysis. An acceleration voltage of 15 kV was used for EDS and EBSD. Samples for SEM-based techniques were prepared using standard metallographic techniques. For light optical microscopy (LOM) and secondary electron (SE) imaging in the SEM, the samples were etched using 5 vol% nital (HNO_3 in ethanol).

APT samples were prepared by the standard lift-out process[26] in a Thermo Fisher Scientific Helios NanoLab 600i dual-beam FIB/SEM device. We sharpened the APT tips by annular milling at 30 kV followed by a low kilovolt milling at 5 kV for 1 min. APT tips from the middle of the precipitation-hardened band as well as from the softer region in between the precipitation-hardened bands were prepared.

APT experiments were performed in a Cameca LEAP 5000 XR and a 5000 XS in laser-pulsing mode. A pulse frequency between 125 and 333 kHz on the 5000 XR and between 250 and 625 kHz on the 5000 XS, a pulse energy between 40 and 75 pJ, and a temperature between 40 and 60 K were used. The detection rate was set between 1 and 4%. The commercial Integrated Visualization and Analysis Software (IVAS, version 3.8.2) was used to reconstruct the tip volume. Voxel-based analysis was performed with a grid spacing of 1 nm and a delocalization of 2 nm.

When analysing the η-phase $(Ni,Fe)_3Ti$ precipitates, one has to take into account the local magnification artefacts due to the differences in evaporation field between the matrix and the precipitate[27–29]. The field of evaporation of the η-phase precipitate is much higher than that of the matrix, which leads to an 'outwards projection' of the ion trajectories of the precipitate phase. Consequently, this leads to an artificially increased apparent volume as well as a lowered apparent density of the precipitate in the APT reconstruction. Simply extracting the volume enclosed by the isocomposition surfaces to calculate precipitate volume would lead to a substantial overestimation of the volume fraction. Therefore, we extracted the number of Ti atoms enclosed by the isocomposition surfaces $N_{Ti,prec}$ and calculated the volume fraction by dividing the number of atoms inside the precipitates (that is, $4 \times N_{Ti,prec}$) by the number of atoms in the entire reconstruction. In addition, we corrected the volume fraction for the slightly larger density of Ni_3Ti compared with Fe.

We acquired multiple datasets from multiple APT tips for each phase in both regions (that is, martensite as well as austenite in the precipitation-hardened bands as well as in the soft regions). None of the measurements of austenite and ferrite in the soft region as well as austenite in the hard region showed any indications of notable clustering of Ti and/or Ni. For the calculation of the precipitate volume fraction in the martensitic phase in the hard regions, we averaged over four individual APT measurements sampling over 24 million nm^3 in total (equivalent to detecting over 1.3 billion ions).

From correlative EBSD and EDS measurements, it is known that the interdendritic regions enriched in Ni and Ti represent the austenite phase, while martensite is depleted in both elements and represents the dendritic regions. Using this knowledge, it is possible to relate each APT reconstruction to either austenite or martensite via the Ti and Ni content in the APT measurements. In the soft regions, the mean Ni and Ti content in the austenite was 21.5 ± 2.7 at% and 8.1 ± 2.9 at% and in the martensite was 16.7 ± 0.3 at% and 3.0 ± 0.5 at%. In the hard regions, the mean Ni and Ti content in the austenite was 19.3 ± 0.1 at% and 6.5 ± 0.7 at% and in the martensite 16.3 ± 0.2 at% and 3.1 ± 0.5 at%.

Vickers hardness measurements along the build direction of the sample were performed using a LECO M-400-G (LECO Instrumente).

Thermo-Calc software (version 2016) together with the TCFE7 database was used to calculate the Gibbs energies of the face-centred cubic (fcc) and body-centred cubic (bcc) phases as a function of the Ti content.

For tensile testing, a Zwick Z100 equipped with a laserXtens 2 HP/TZ laser extensometer was used. Tests were performed at room temperature at a strain rate of 10^{-3} mm min^{-1} on dog-bone-shaped samples with a gauge length of 25 mm, a thickness of 1 mm, a gauge width of 5 mm and a total length of 45 mm. These specimens were machined with the gauge length parallel to the laser scan direction. Further, smaller, test specimens (4-mm gauge length, 2-mm gauge width and 0.35-mm gauge thickness) were machined with the gauge length parallel to the build direction (that is, perpendicular to the layered structure) and the gauge width parallel to the laser scan direction. These samples were tested in

a Kammrath and Weiss tensile testing stage at room temperature and at a strain rate of 2 μm s^{-1}. The strain was measured by digital image correlation using the Aramis software (GOM). The results are shown in the Extended Data Fig. 9b.

Dilatometer experiments were carried out in a Bähr Thermoanalyse DIL805A/D dilatometer using hollow cylindrical specimens with a height of 1 cm, an outer diameter of 4.5 mm and an inner diameter of 2 mm at a heating rate of 600 °C min^{-1} and a cooling rate of 160 °C min^{-1}.

Supplementary discussion of the alloy design concept

LAM is currently mostly applied to conventional alloys, with compositions not optimized for the specific conditions encountered during the fabrication. This can lead to severe problems regarding processability and furthermore leaves aside opportunities for alloy design and tailored microstructures. Here we designed a new steel optimized and tailor-made for LAM exploiting two specific conditions of LAM, namely rapid quenching and cyclic re-heating (the so-called IHT). We considered three key requirements that the steel needed to fulfil:

(1) A martensitic microstructure after fabrication.

(2) An M_s value that lies in a control window that is readily accessible to the digital control exerted during the DED process

(3) A kinetic window to respond in the desired way and quickly to the IHT with a substantial precipitation reaction.

All three factors drove the design of the steel, that is, the selection of the composition. The exact precipitation kinetics (especially during the very nonlinear heat treatment of the IHT) as well as the microstructure after rapid quenching during LAM turned out to be challenging to predict. Therefore, we used a rapid alloy prototyping approach, whereby we built a compositionally graded sample using DED. This preliminary fabrication allowed us to efficiently screen the microstructure and the response to the IHT as a function of the alloy composition. For an Fe–Ni–Al steel, this approach has been outlined and explained in more detail in our previous study[9].

Data availability

The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information and extended data files. Raw data are available from the corresponding author upon reasonable request.

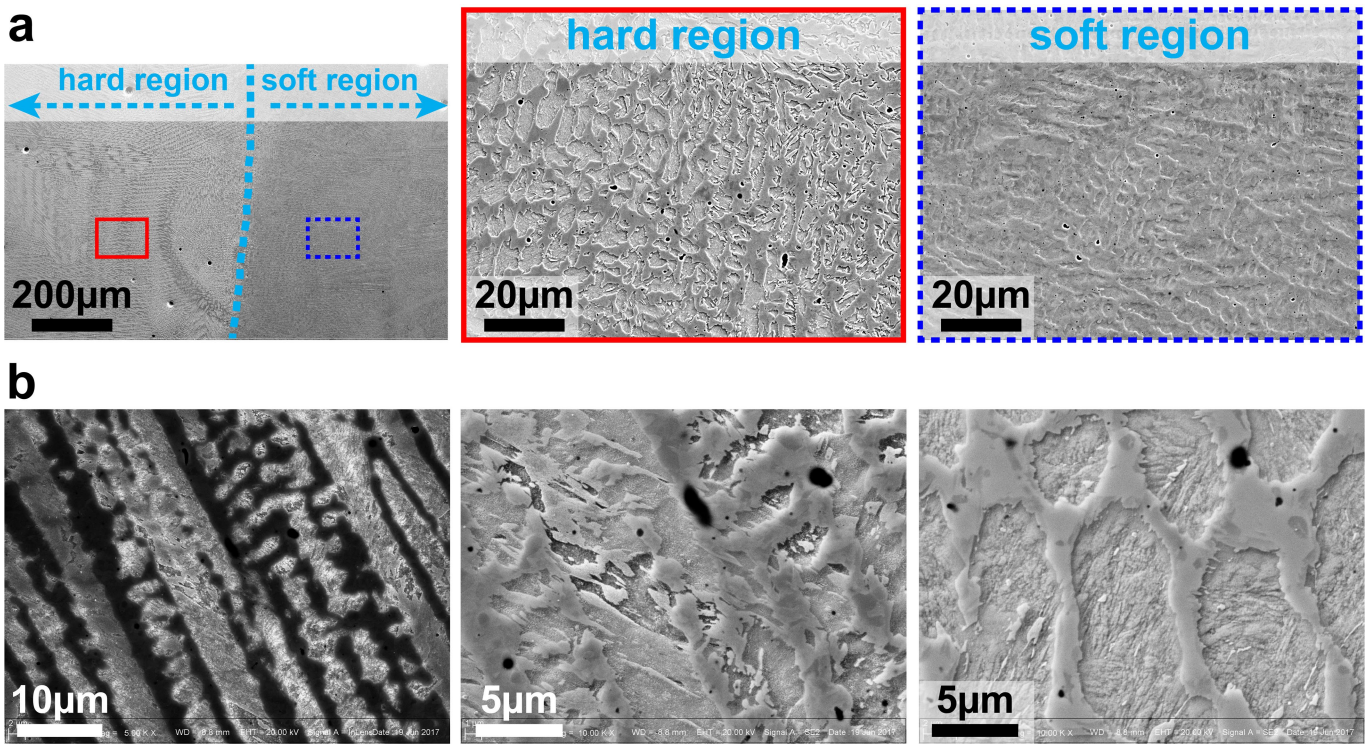
26. Larson, D. J., Prosa, T. J., Ulfig, R. M., Geiser, B. P. & Kelly, T. F. *Local Electrode Atom Probe Tomography: A User's Guide* (Springer, 2013).
27. Vurpillot, F., Bostel, A. & Blavette, D. Trajectory overlaps and local magnification in three-dimensional atom probe. *Appl. Phys. Lett.* **76**, 3127–3129 (2000).
28. Lefebvre, W. et al. 3DAP measurements of Al content in different types of precipitates in aluminium alloys. *Surf. Interface Anal.* **39**, 206–212 (2007).
29. Marquis, E. A. & Vurpillot, F. Chromatic aberrations in the field evaporation behavior of small precipitates. *Microsc. Microanal.* **14**, 561–570 (2008).
30. Hellman, O. C., Vandenbroucke, J. A., Rüsing, J., Isheim, D. & Seidman, D. N. Analysis of three-dimensional atom-probe data by the proximity histogram. *Microsc. Microanal.* **6**, 437–444 (2000).
31. Casati, R., Lemke, J. & Vedani, M. Microstructure and fracture behavior of 316L austenitic stainless steel produced by selective laser melting. *J. Mater. Sci. Technol.* **32**, 738–744 (2016).
32. Kim, H., Liu, Z., Cong, W. & Zhang, H. C. Tensile fracture behavior and failure mechanism of additively-manufactured AISI 4140 low alloy steel by laser engineered net shaping. *Materials* **10**, 1283 (2017).
33. Kasper, R. & Faul, H. Charpy-V subsize specimens: measurements of steel impact properties. *Materialpruefung* **43**, 18–21 (2001).
34. Lucas, G. E., Odette, G. R., Sheckherd, J. W. & Krishnadev, M. R. Recent progress in subsized Charpy impact specimen testing for fusion reactor materials development. *Fusion Technol.* **10**, 728–733 (1986).
35. Schill, R., Forget, P. & Sainte Catherine, C. Correlation between Charpy-V and sub-size Charpy tests results for an un-irradiated low alloy RPV ferritic steel. In *Thirteenth European Conference on Fracture* (Elsevier, 2000).
36. Kempen, K., Yasa, E., Thijs, L., Kruth, J. P. & Van Humbeeck, J. Microstructure and mechanical properties of selective laser melted 18Ni-300 steel. *Phys. Proc.* **12**, 255–263 (2011).

Acknowledgements We are grateful to U. Tezins and A. Sturm for their support to the FIB and APT facilities at MPIE, to H. Faul and A. Jansen for their help with tensile tests, and to M. Adamek for his help with dilatometer experiments. A. Kwiatkowski da Silva and P. Bajaj are acknowledged for their input and discussions regarding thermodynamics and additive manufacturing respectively. We thank C. Brunner-Schwer for his support in conducting the DED experiments.

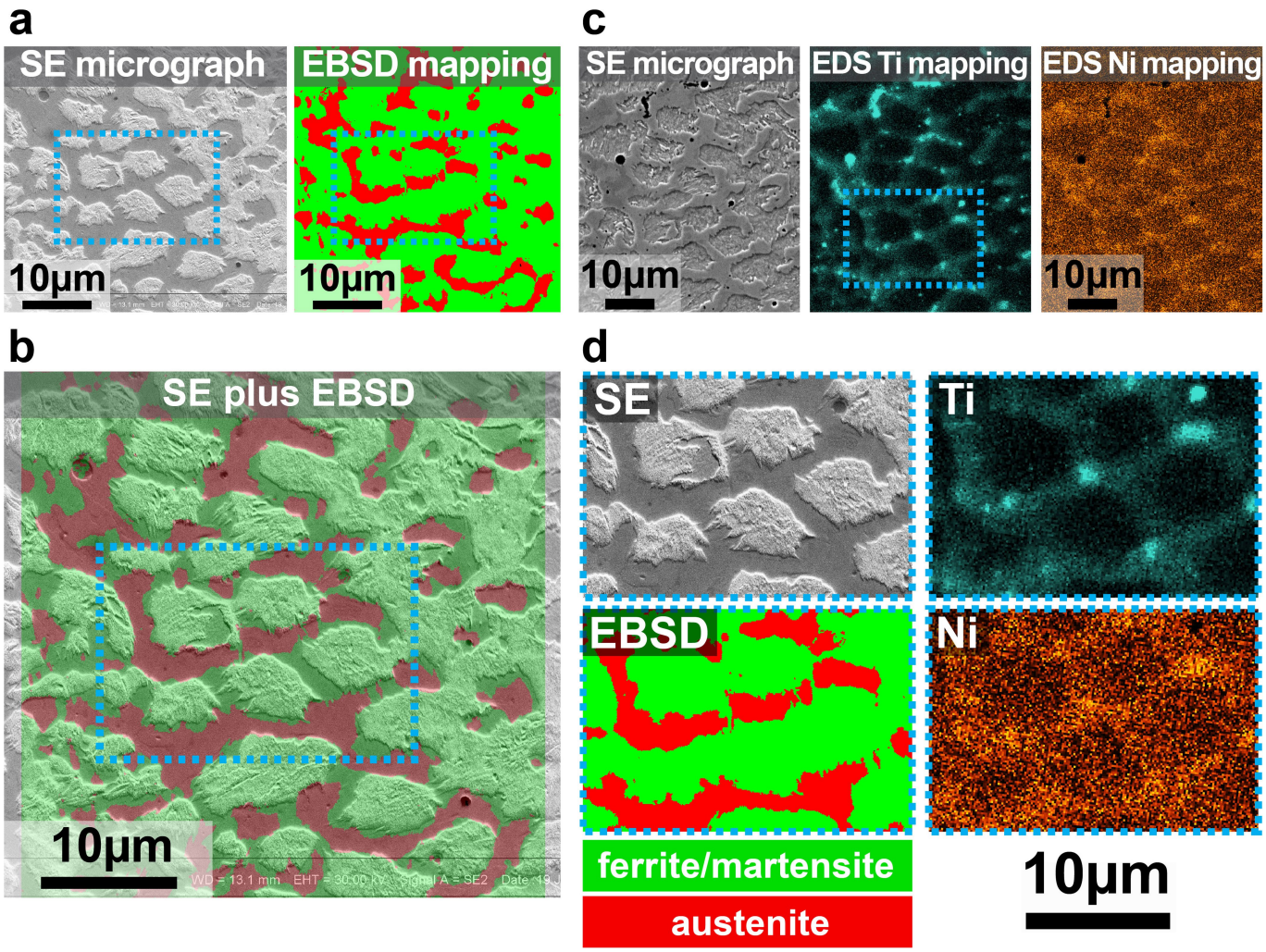
Author contributions P.K. performed the microstructure analysis and corresponding data analysis including EDS, EBSD, FIB and APT and the analysis of dilatometer experiments and tensile tests. M.B.W. produced all samples used in this study by DED and acquired the experimental thermal profiles as well as the optical micrographs. E.A.J., A.W., B.G. and D.R. designed the study and acquired funding. P.K. wrote the initial draft. All authors contributed to reviewing and editing the manuscript and discussing and interpreting all the results.

Competing interests The authors declare no competing interests.

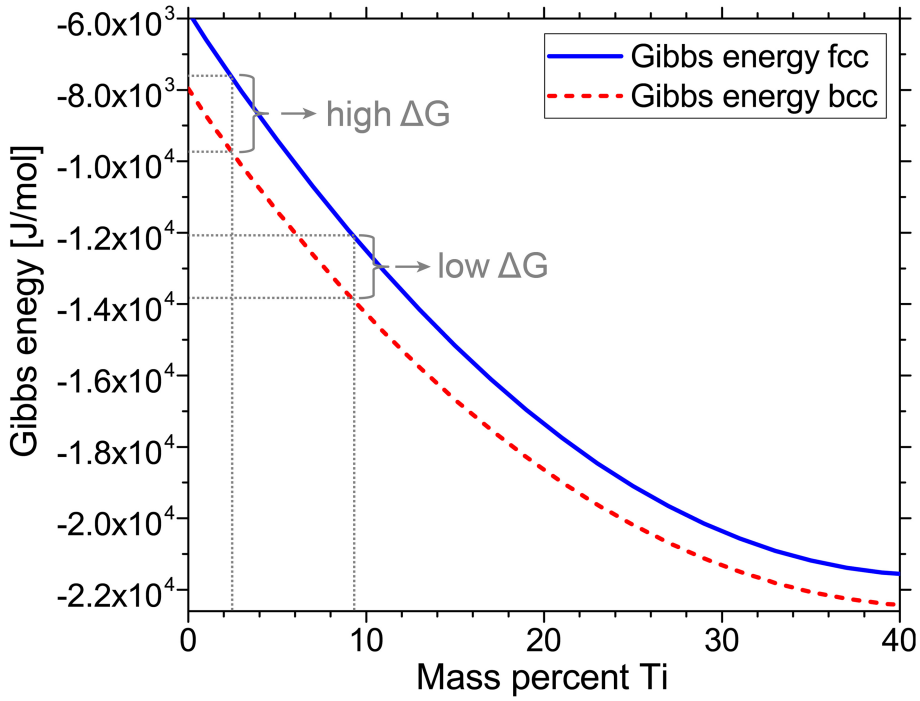
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2409-3.
Correspondence and requests for materials should be addressed to P.K.
Peer review information *Nature* thanks Claire Davis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



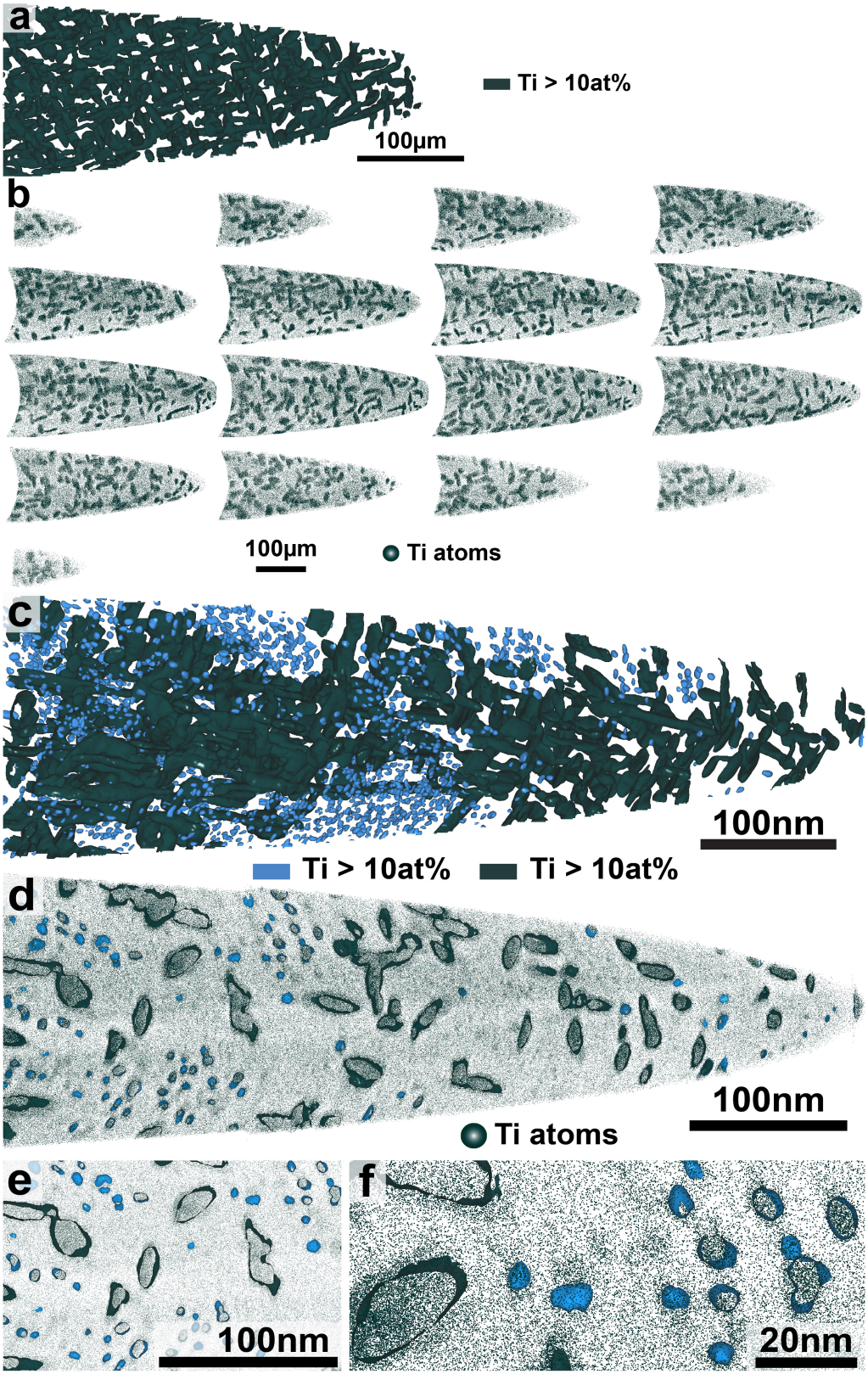
Extended Data Fig. 1 | Electron micrographs of the soft and hard regions. In the SE micrographs, the hard region appears bright due to the rougher surface emitting more SEs, while the soft region appears darker due to the smooth surface (that is, opposite of how these two regions appear in optical micrographs). **a**, The interface between the soft and hard regions as well as a zoom to both regions. It is apparent that in the hard region there are two distinct phases: one with a rough surface, which is the martensite with $(Ni,Fe)_3Ti$ precipitates, and one with a smooth surface, which is austenite (and does not contain any precipitates). In the soft region, both austenite and martensite have a smooth surface as the martensite does not contain any precipitates in the soft regions. **b**, Further examples of the rough surfaces in hard regions at higher magnifications. The sample was slightly etched with 5 vol% nital for 10 s at room temperature.



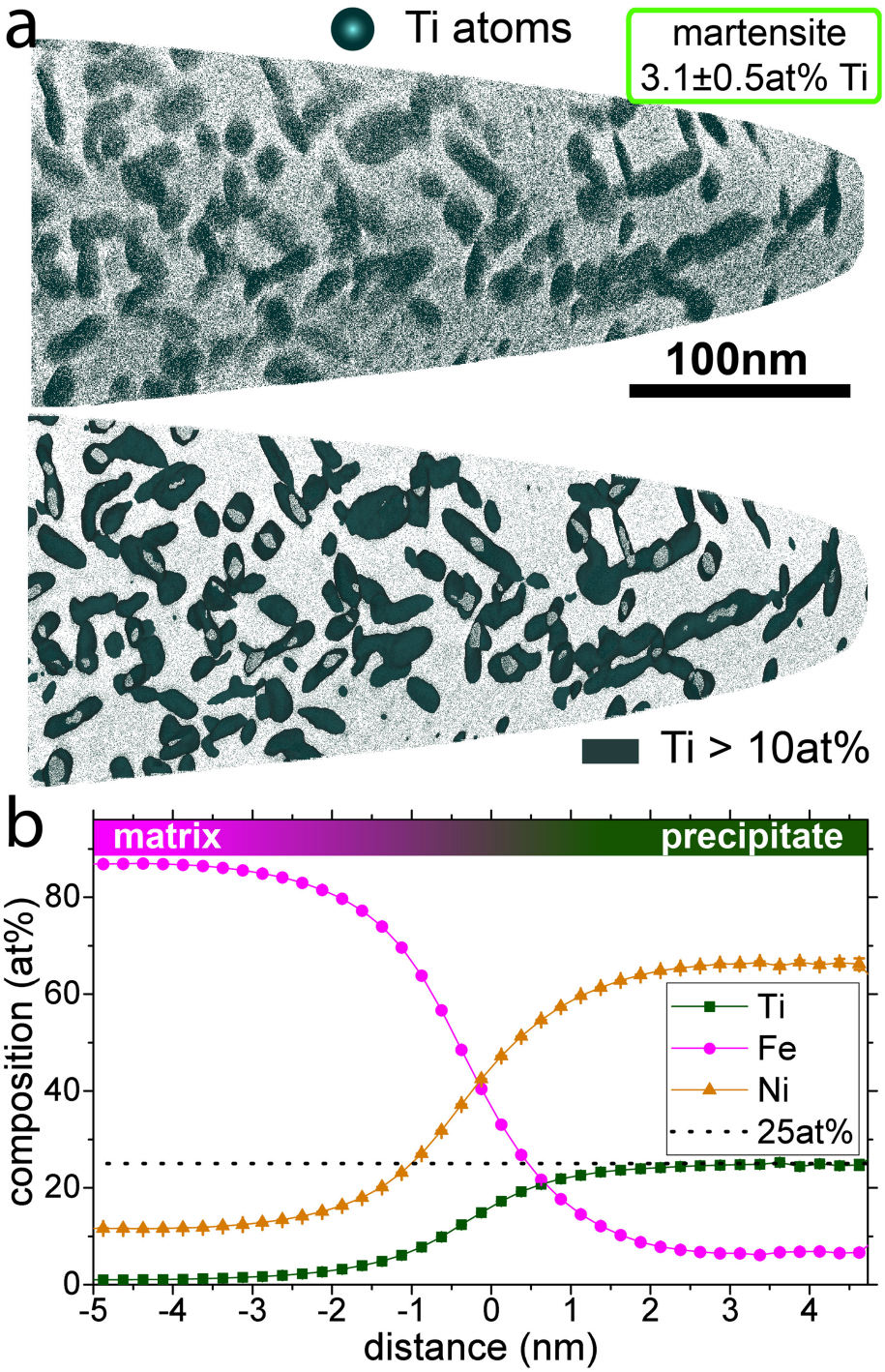
Extended Data Fig. 2 | Microstructure characterization. High-resolution EBSD together with correlative EDS elemental mapping of a location within a hard region, showing that the austenite is enriched in Ti and Ni. **a**, SE micrograph (left) and EBSD mapping (right) of the same location. **b**, Overlapping the SE micrograph and the EBSD phase map from **a** show that austenite has a smooth surface and appears darker than martensite with a rough surface. **c**, The EDS element mapping shows that the smooth regions (that is, austenite) are enriched in Ti and Ni. These regions represent the interdendritic regions. **d**, EDS and EBSD are brought together: the area shown in **d** is marked by light blue dashed boxes in **a**–**c**. It becomes apparent that austenite has a smooth surface and is enriched in Ti and Ni, whereas martensite has a rough surface and is depleted in Ti and Ni.



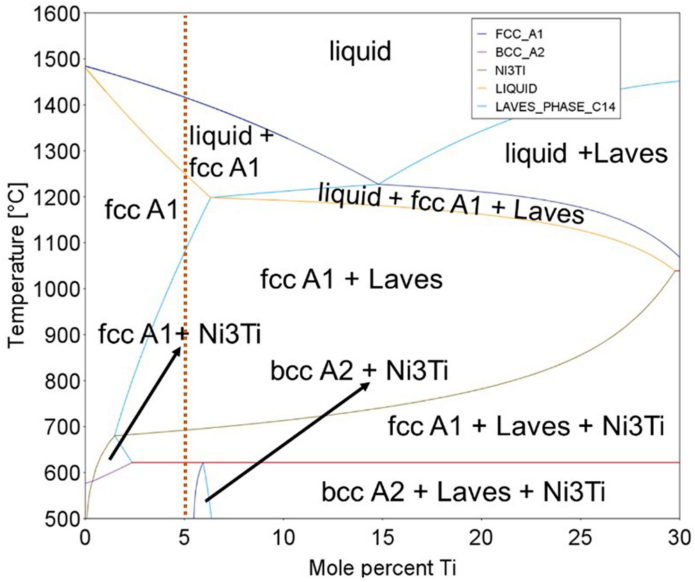
Extended Data Fig. 3 | Thermodynamic calculation of the driving force for martensite formation. The Gibbs free energies of single-phase bcc and fcc Fe–Ni–Ti at room temperature for a variable Ti content and a fixed Ni content of 19 wt%. It is apparent that there is a higher energy difference between the fcc and bcc structures (that is, a higher driving force for martensite formation) at lower Ti contents than at higher Ti contents. The two Ti compositions highlighted in the graph are 2.3 wt% Ti and 8.6 wt% Ti, which are typical compositions for the martensite/dendritic region and the austenite/interdendritic region, respectively, and for which the driving force for martensite formation is −2,100 J mol^{-1} and −1,780 J mol^{-1}, respectively.



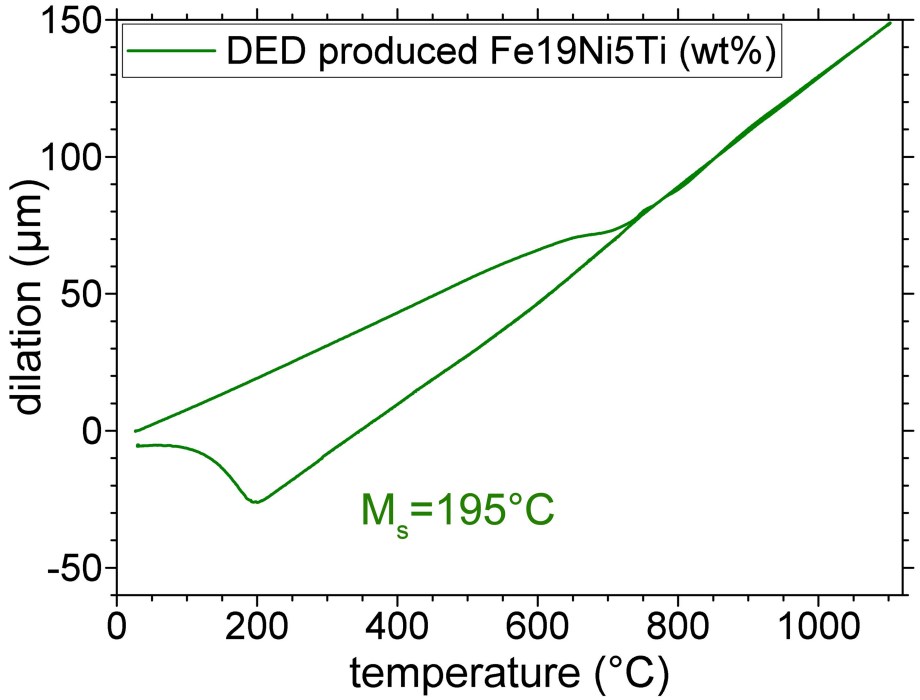
Extended Data Fig. 4 | Serial sections through an APT reconstruction. **a**, The $(Ni,Fe)_3Ti$ precipitates are marked by a 10 at% Ti isocomposition surface in the APT reconstruction. **b**, Consecutive slices through the same dataset from one side of the tip to the other. Each slice is 10 nm thick and all Ti atoms within this slice are shown. This sequence of images illustrates the complex shape and morphology and three-dimensional arrangement of the network of η-phase $(Ni,Fe)_3Ti$ precipitates created by IHT during the DED process. **c**–**f**, Reconstruction of an APT volume that contains small spherical precipitates a few nanometres in diameter in addition to the plate-shaped interconnected network of precipitates. Both precipitate types are η-phase $(Ni,Fe)_3Ti$ and are marked by means of 10 at% Ti isocomposition surfaces. The plate-shaped network is depicted in dark green and the small spherical precipitates are depicted in light blue. In **c**, the whole dataset is shown, and in **d**–**f**, only a thin slice of 5 nm thickness is shown. In addition to the isocomposition surfaces, Ti atoms are shown in **d**–**f**. Panels **e** and **f** are enlarged sections of the image shown in **c**.



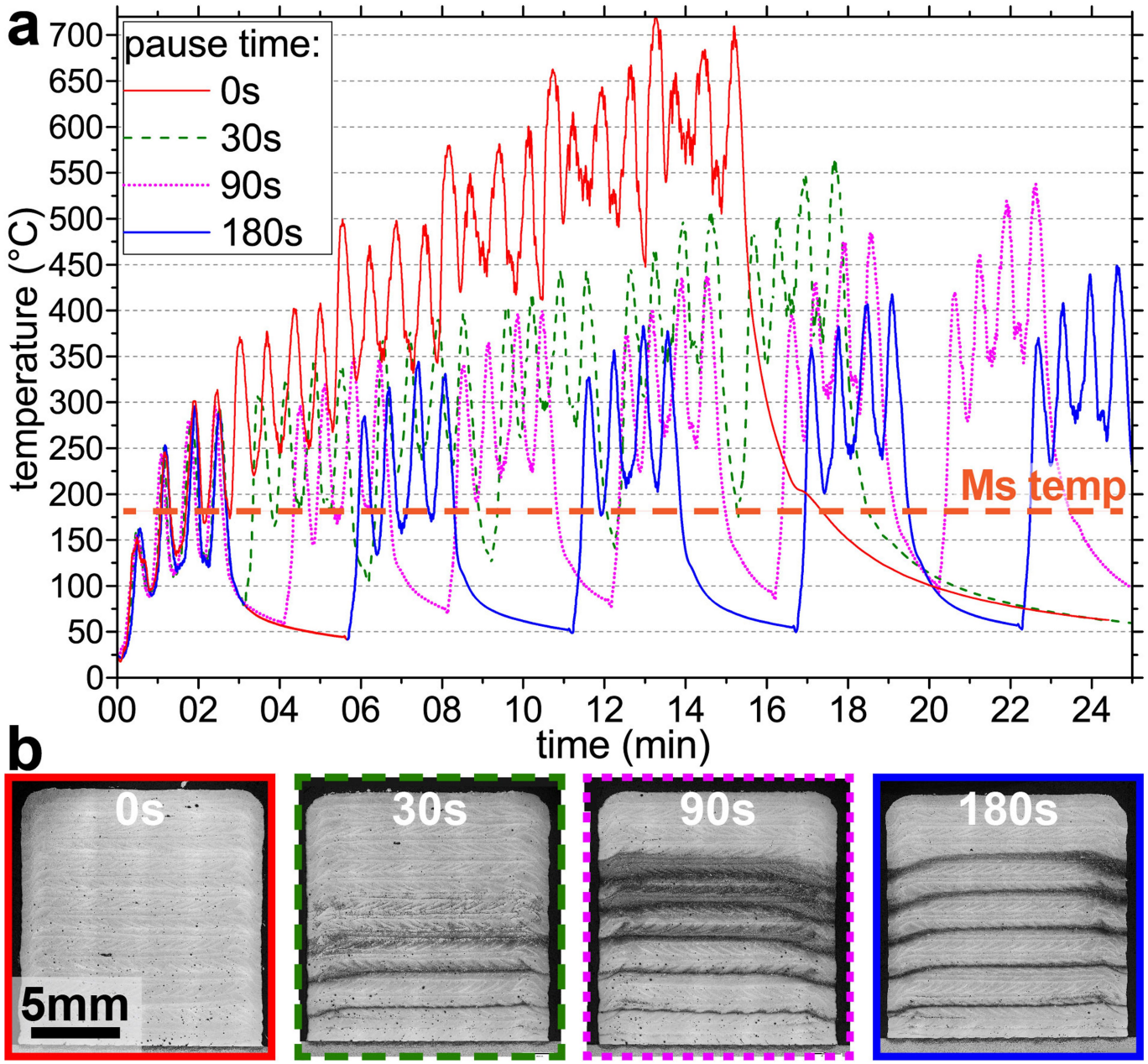
Extended Data Fig. 5 | Composition of the η-phase precipitates. a, A Ti atom map of a 5-nm-thick slice through the reconstructed volume from the martensitic phase in the hard region. The top part shows the atom map only; in the bottom part, precipitates are also highlighted by a set of isocomposition surfaces encompassing regions containing more than 10 at% Ti (dark green). **b**, A proximity histogram, that is, composition profile as a function of the distance to this isocomposition surface[30], calculated for all imaged precipitates in the dataset. The average Ti content fits the expected 25 at% almost perfectly. Fe replaces some of the Ni from the prototype Ni_3Ti phase, rendering it a $(Ni,Fe)_3Ti$ phase with approximately 6–7 at% Fe and 66–67 at% Ni.



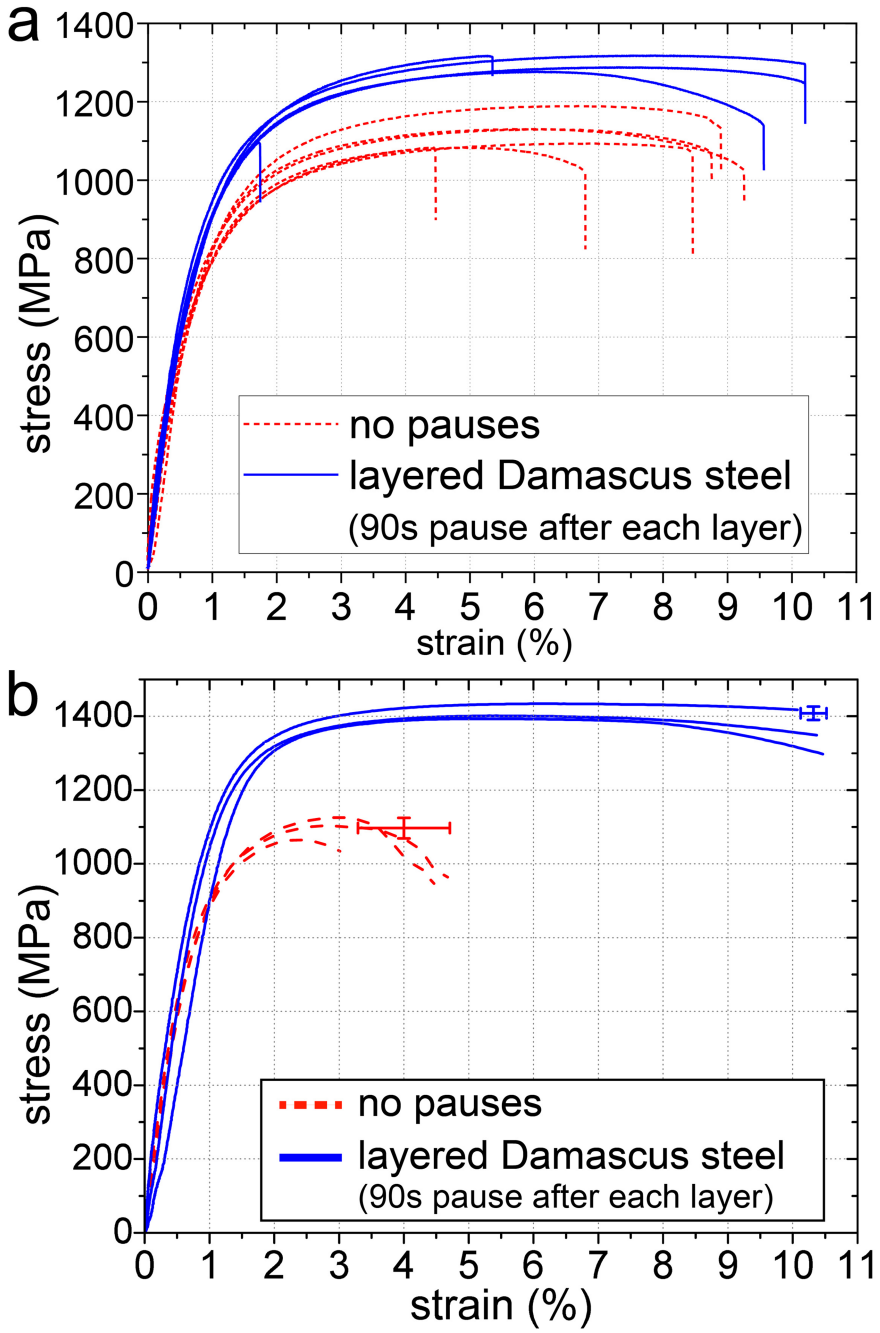
Extended Data Fig. 6 | Pseudobinary phase diagram for the Fe19Ni-*x*Ti (at%) alloy. The phase diagram was calculated using the Thermo-Calc software in conjunction with the TCFE7 database. The dashed line at 5 at% Ti highlights the phases that can be expected for the Fe19Ni5Ti (at%) steel used in this study: liquid, fcc A1 austenite, η-phase Ni_3Ti, Laves phase and bcc A2 ferrite/martensite.



Extended Data Fig. 7 | Determination of M_s. A dilatometer curve acquired on a DED-produced Fe19Ni5Ti (wt%) sample (with no pause) is depicted. The double tangent method was used to determine M_s as 195 °C. Further dilatometer specimen from the same DED sample as well as DED samples with a 90-s pause time were measured. On all measurements, the martensite start temperatures are within 10 °C.

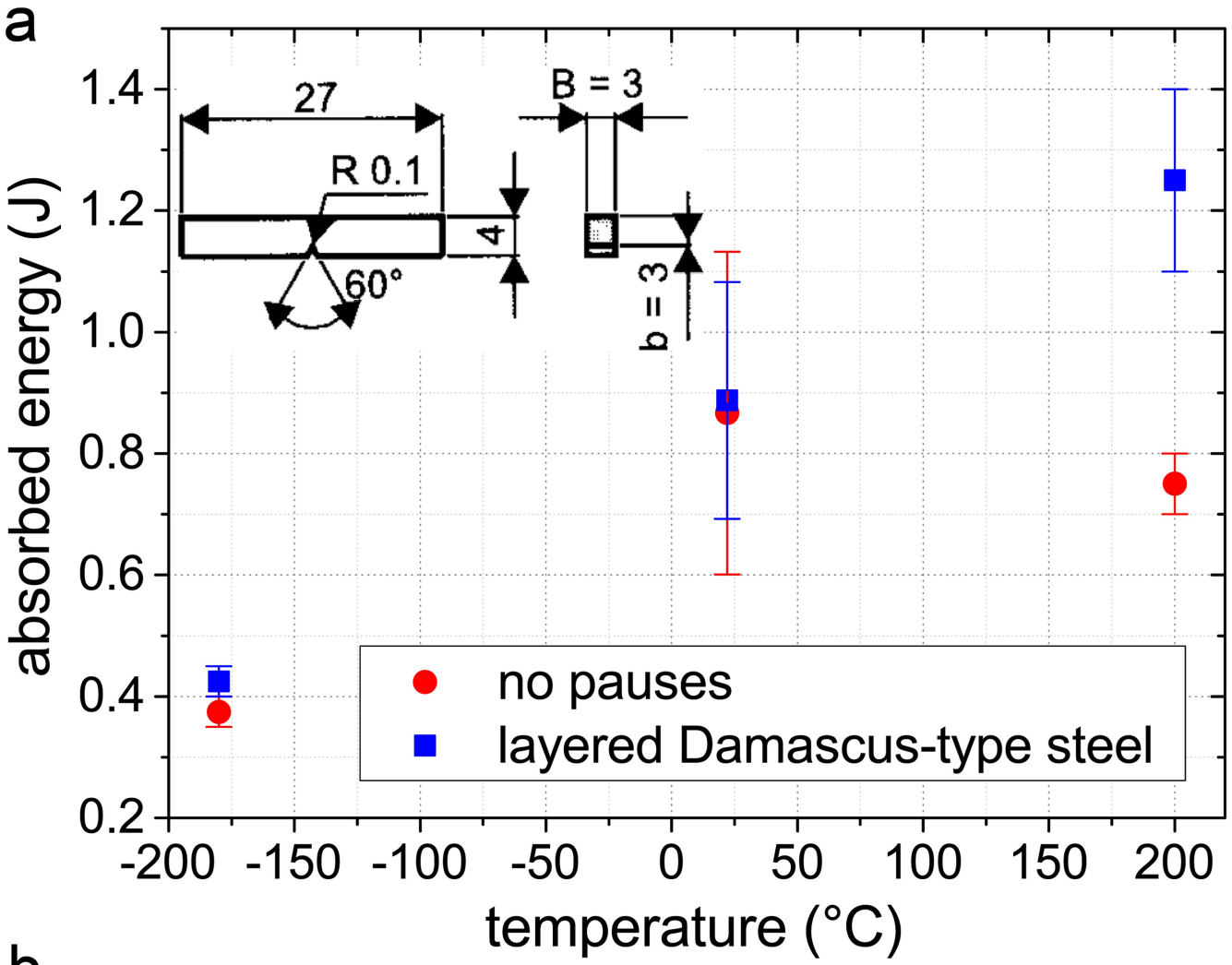


Extended Data Fig. 8 | Experimental time–temperature profiles. **a**, Experimental time–temperature profiles acquired with a pyrometer on the surface of the sample during the DED build at different pause times after each fourth layer. It becomes apparent that without pauses, the temperature increases during the entire fabrication and only drops notably when a pause time is introduced. The dashed orange line corresponds to M_s. **b**, Optical micrographs of the samples built with the corresponding pause times.



Extended Data Fig. 9 | Tensile curves. a, The testing direction is parallel to the laser scan direction. Tensile tests show a substantial improvement in strength as well as ductility when a pause is introduced in the manufacturing process. The pause allows the material to partially transform to martensite and then allows the IHT to trigger $(Fe,Ni)_3Ti$ precipitates in the martensite. The results show a few tensile specimens that fracture prematurely at low strains, which is due to additive-manufacturing-process-related defects[31,32]. These outliers rather represent the additive manufacturing process and show that there is potential for future process optimization. The specimens containing fewer defects and therefore higher strength and ductility show the actual potential of the newly designed maraging steel. In Fig. 5, we show two representative curves for each condition. For the condition '90-s pause each layer', we omitted the one sample fracturing prematurely at 1.7% strain as well as the samples with the highest strength and lowest strength and show the two curves in between. For the condition 'no pauses', we omitted the two samples fracturing at the lowest strains of 4.5% and 6.7% as well as the samples with the lowest strength and highest strength. **b**, The testing direction is parallel to the build direction. Owing to limitations in the size of the DED-produced samples, we used small tensile specimens with a gauge length of 4 mm, a width of 2 mm and a thickness of 0.35 mm to test the tensile properties along the build direction (that is, perpendicular to the layered structure). The tensile specimens were machined with the gauge width parallel to the laser scan direction. There is a notable increase in strength and ductility due to the layered, Damascus-type structure. However, due to the smaller size, compared with the tensile specimen machined along the laser scan direction, a direct comparison between the two is difficult. Both, the Damascus-type layered steel as well as the one that was produced without pauses in between layers show higher strengths along the build direction than in the laser scan direction. While this could be due to the anisotropy of the material, the smaller tensile specimen geometry might also have a role.

a



b

	Temperature [°C]	Absorbed energy [J]	Normalized absorbed energy (factor B*b) [J]	Normalized absorbed energy (factor B*b²) [J]
No pause	-180	0.38	3.33	8.89
	22	0.87	7.70	20.54
	200	0.75	6.67	17.78
Damascus-type	-180	0.43	3.78	10.07
	22	0.89	7.89	21.04
	200	1.25	11.11	29.63

Extended Data Fig. 10 | Impact toughness. a, The absorbed energies of subsized V-notch Charpy specimens at three different temperatures of −180 °C, 22 °C and 200 °C. The inset shows the geometry of the used subsized specimens. Charpy specimens were machined along the laser scan direction of the DED sample with the *B* direction normal to the layers and the *b* direction parallel to the layers. The values shown in the graph are an average of three specimens at 22 °C and two specimens at −180 °C and 200 °C. **b**, The values of the absorbed energy in joules in the Charpy V-notch impact testing carried out on subsized specimens shown in the inset in **a**. Two different normalizing factors are used to convert the results of the subsized specimen to standard specimen ($55 \times 10 \times 8\ mm^3$): the fracture area $B \times b$ and the fracture volume $B \times b^2$ (see, for example, refs. [33–35]). It is noted that such normalizing factors are material dependent and there is no literature available on the selection of normalizing factors for additively manufactured maraging steels. The converted values presented in this table should therefore only be regarded as a rough estimate of the impact toughness on standard samples. Nevertheless, the Fe19Ni5Ti (wt%) samples investigated in this study show a high impact toughness compared with 4.9 J (standard V-notch samples) of laser-powder-bed-fusion-produced 18Ni-300 maraging steel in the aged condition (5 h at 480 °C)[36].

Mapping the emergence of molecular vibrations mediating bond formation

https://doi.org/10.1038/s41586-020-2417-3

Received: 18 October 2019

Accepted: 16 April 2020

Published online: 24 June 2020

Check for updates

Jong Goo Kim[1,2,3], Shunsuke Nozawa[4,5], Hanui Kim[1,2,3], Eun Hyuk Choi[1,2,3], Tokushi Sato[6,7], Tae Wu Kim[1,2,3], Kyung Hwan Kim[8], Hosung Ki[1,2,3], Jungmin Kim[1,2,3], Minseo Choi[1,2,3], Yunbeom Lee[1,2,3], Jun Heo[1,2,3], Key Young Oang[9], Kouhei Ichiyanagi[4], Ryo Fukaya[4], Jae Hyuk Lee[10], Jaeku Park[10], Intae Eom[10], Sae Hwan Chun[10], Sunam Kim[10], Minseok Kim[10], Tetsuo Katayama[11,12], Tadashi Togashi[11,12], Sigeki Owada[11,12], Makina Yabashi[11,12], Sang Jin Lee[1,2,3], Seonggon Lee[1,2,3], Chi Woo Ahn[1,2,3], Doo-Sik Ahn[1,2,3], Jiwon Moon[13], Seungjoo Choi[14], Joonghan Kim[13], Taiha Joo[8], Jeongho Kim[14], Shin-ichi Adachi[4,5] & Hyotcherl Ihee[1,2,3]✉

Fundamental studies of chemical reactions often consider the molecular dynamics along a reaction coordinate using a calculated or suggested potential energy surface[1–5]. But fully mapping such dynamics experimentally, by following all nuclear motions in a time-resolved manner—that is, the motions of wavepackets—is challenging and has not yet been realized even for the simple stereotypical bimolecular reaction[6–8]: A–B + C → A + B–C. Here we track the trajectories of these vibrational wavepackets during photoinduced bond formation of the gold trimer complex $[Au(CN)_2^-]_3$ in an aqueous monomer solution, using femtosecond X-ray liquidography[9–12] with X-ray free-electron lasers[13,14]. In the complex, which forms when three monomers A, B and C cluster together through non-covalent interactions[15,16], the distance between A and B is shorter than that between B and C. Tracking the wavepacket in three-dimensional nuclear coordinates reveals that within the first 60 femtoseconds after photoexcitation, a covalent bond forms between A and B to give A–B + C. The second covalent bond, between B and C, subsequently forms within 360 femtoseconds to give a linear and covalently bonded trimer complex A–B–C. The trimer exhibits harmonic vibrations that we map and unambiguously assign to specific normal modes using only the experimental data. In principle, more intense X-rays could visualize the motion not only of highly scattering atoms such as gold but also of lighter atoms such as carbon and nitrogen, which will open the door to the direct tracking of the atomic motions involved in many chemical reactions.

The $[Au(CN)_2^-]_3$ complex serves as a valuable model system for studying photoinitiated processes in solution. Irradiation with ultraviolet light excites $[Au(CN)_2^-]_3$ from its ground state (S_0) to the singlet state (S_1), which within 20 fs undergoes intersystem crossing to reach a triplet excited state (T_1')[17]. A further transition from T_1' to another triplet excited state (T_1) then occurs with a time constant of about 1–2 ps, completing the formation of covalent bonds and transformation of the complex from a bent to a linear structure[9,17,18] (see Supplementary Information for details of the notations of electronic states).

Formation of the bonds could involve any of the three possible candidate trajectories sketched in Fig. 1b. The equilibrium structure in the ground state determines the position of the Franck–Condon (FC) region in the excited state; the excited-state wavepacket created in the FC region can be considered as the reactants (A + B + C) of the reaction. This wavepacket moves towards the equilibrium structure of T_1', which is the product (A–B–C) with two equivalent covalent Au–Au bonds. Using three-dimensional nuclear coordinates R_{AB}, R_{BC} and R_{AC}, if the FC region is located at the point at which R_{AB} is shorter than R_{BC}, the shortest pathway connecting the FC region and the equilibrium structure of T_1' is path 2, corresponding to the concerted bond formation. Alternatively, two covalent bonds can form sequentially in time (that is, asynchronously), as in path 1 and path 3, which differ only by the order in which covalent bonds are formed: path 1 represents a pathway in which the covalent bond between A and B is formed first, and path 3 represents the case in which the bond between B and C is formed first (Fig. 1c). To determine the position of the FC region and whether the reaction trajectory involves concerted or asynchronous bond formation, the initial motions of the wavepacket starting from the FC region

[1]Department of Chemistry, KAIST, Daejeon, Republic of Korea. [2]KI for the BioCentury, KAIST, Daejeon, Republic of Korea. [3]Center for Nanomaterials and Chemical Reactions, Institute for Basic Science (IBS), Daejeon, Republic of Korea. [4]Photon Factory, Institute of Materials Structure Science, High Energy Accelerator Research Organization (KEK), Tsukuba, Japan. [5]Department of Materials Structure Science, School of High Energy Accelerator Science, The Graduate University for Advanced Studies, Tsukuba, Japan. [6]Center for Free-Electron Laser Science (CFEL), Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany. [7]European XFEL, Schenefeld, Germany. [8]Department of Chemistry, Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea. [9]Radiation Center for Ultrafast Science, Quantum Optics Division, Korea Atomic Energy Research Institute (KAERI), Daejeon, Republic of Korea. [10]Pohang Accelerator Laboratory, Pohang, Republic of Korea. [11]Japan Synchrotron Radiation Research Institute (JASRI), Sayo, Japan. [12]RIKEN SPring-8 Center, Sayo, Japan. [13]Department of Chemistry, The Catholic University of Korea, Bucheon, Republic of Korea. [14]Department of Chemistry, Inha University, Incheon, Republic of Korea. ✉e-mail: hyotcherl.ihee@kaist.ac.kr

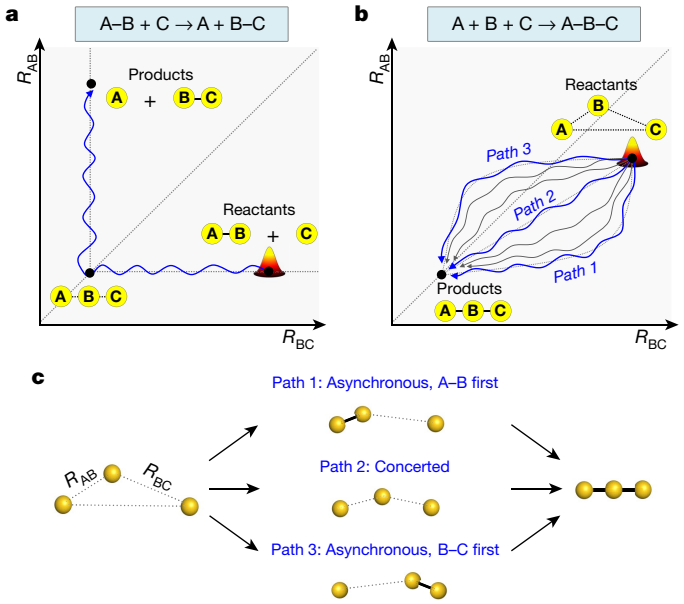


Fig. 1 | Schematics of the mechanisms for reactions involving three atoms and two bonds. a, A representative reaction trajectory for A–B + C → A + B–C. **b**, Representative reaction trajectories for A + B + C → A–B–C. **c**, Candidate pathways of the reaction in **b**. Path 2 represents a pathway whereby the two covalent Au–Au bonds are formed simultaneously, corresponding to a concerted bond formation mechanism. Path 1 and path 3 represent pathways whereby the two bonds are formed sequentially in time, corresponding to an asynchronous bond-formation mechanism. Path 1 and path 3 are distinct, depending on which bond is formed first, as described in the text. To determine the reaction pathway, the initial motion of the wavepacket must be tracked.

on a multidimensional potential energy surface (PES) need to be observed directly—something not achieved in previous studies (see Extended Data Fig. 1 for details).

Femtosecond electron and X-ray scattering[10,11,19–26] and X-ray absorption spectroscopy[19,27,28] have both the structural sensitivity and the temporal resolution needed for probing ultrafast changes of molecular structure in real space and real time. They have been used to observe vibrational motions[10,11,19–24,27,28], but mostly to observe diatomic molecules that have only a single nuclear coordinate or polyatomic molecules that were approximated as pseudo-diatomic species. This reflects the challenging nature of tracking wavepacket motions in the multidimensional nuclear coordinates of polyatomic molecules.

We accomplished this task using time-resolved X-ray liquidography (TRXL)[9–12], also known as time-resolved X-ray solution scattering. Time-resolved difference scattering curves, $q\Delta S(q, t)$, for the momentum-transfer vector $q = (4\pi/\lambda)\sin(2\theta/2)$, where λ is the X-ray wavelength and 2θ is the scattering angle, and a measurement time t, are shown in Extended Data Fig. 2a. Details of experimental procedures and data analysis are described in Methods and Supplementary Information. The temporal changes of $q\Delta S(q,t)$ can be determined from the first two right singular vectors (RSVs) obtained from a singular value decomposition (SVD) of $q\Delta S(q, t)$. The two RSVs are well fitted by an exponential function with a time constant of 1.1 ± 0.1 ps (mean ± s.e.m.), which is related to the T_1'-to-T_1 transition[9], convoluted with the instrument response function, IRF(t) (Extended Data Fig. 2c). Apart from these population kinetics, oscillations are observed in the first to fourth RSVs (Extended Data Fig. 2d). To analyse the oscillations in more detail, we extracted the oscillating components from the experimental $q\Delta S(q,t)$ by subtracting the contributions of the T_1'-to-T_1 transition and the solvent heating, yielding residual difference scattering curves $q\Delta S_{residual}(q,t)$. The two-dimensional $q\Delta S_{residual}(q,t)$ curve in the q domain and the t domain provides direct information on the time-dependent molecular structure and eventually enables us to track the motions of the wavepacket in multidimensional nuclear coordinates.

Figure 2a shows $q\Delta S_{residual}(q,t)$ measured at time delays from −1,040 fs to 2,235 fs. The TRXL signal is sensitive to wavepacket motions in any of the structurally distinct states (that is, the S_0, T_1' and T_1 states) and we therefore first examined which state is associated with the observed features of the residual difference scattering curve. As detailed in Methods and Supplementary Information, the best fits shown in Fig. 2a were obtained by considering the ground state S_0 and excited state T_1', indicating that the residual difference scattering curves arise from wavepacket motions on the PESs of both S_0 and T_1'. In Fig. 2b–e, we show the time-dependent changes of the structural parameters, R_{AB}, R_{BC}, R_{AC} and the Au–Au–Au angle, θ, obtained from the structural analysis. From the time evolution of these structural parameters, the trajectories of the excited-state (Fig. 3a) and ground-state (Fig. 3c) wavepackets can be reconstructed in multidimensional nuclear coordinates, R_{AB} versus R_{BC} versus θ, which describe the relative positions of all three Au atoms in the gold trimer complex. We note that these trajectories are obtained purely on the basis of the experimental data, without recourse to theoretical calculation, thanks to the structural sensitivity of TRXL.

The trajectories of the wavepackets occur in two distinct time regimes: (1) the initial motion on the PES of T_1' starting from the FC region of S_1 at earlier times (t < 360 fs), and (2) subsequent harmonic oscillations around the equilibrium structures of T_1' at later times (t > 360 fs). At earlier times, the excited-state and ground-state wavepackets each move on their own PES to approach their own equilibrium structure. To examine the reaction mechanism of the bond formation, we first inspected the initial motion of the excited-state wavepacket with respect to the progress of the covalent-bond formation and the bent-to-linear transformation. Specifically, as shown in Fig. 3a, the excited-state wavepacket is generated in the FC region (R_{AB} = 3.13 Å, R_{BC} = 3.38 Å, θ = 119°) by an interaction with the pump pulse and then moves on the PES of T_1' towards the equilibrium structure of T_1' (R_{AB} = 2.82 Å, R_{BC} = 2.82 Å, θ = 180°). Along the coordinates of θ, the excited-state wavepacket in T_1' starts from the FC region (θ = 119°) and reaches the equilibrium of T_1' (θ = 180°) within 335 fs, giving the timescale of the bent-to-linear transformation. The progress of the covalent-bond formation can be visualized more clearly by projecting the trajectory of the excited-state wavepacket onto the R_{AB}–R_{BC} plane as shown in Fig. 3a and Supplementary Fig. 5a. The trajectory of the excited-state wavepacket reveals that the formation of two covalent bonds does not occur in a concerted, synchronous manner (as exemplified by path 2 in Fig. 1). Instead, R_{AB} decreases rapidly down to the covalent Au–Au bond length of the equilibrium of T_1' (2.82 Å) at a 35-fs time delay, and at 60 fs it becomes even shorter, reaching the minimum length along the entire trajectory, whereas R_{BC} remains much longer than the covalent bond length (2.82 Å) at those time delays (Fig. 3a and Supplementary Fig. 5a). This trajectory at earlier times indicates that the shape of the PES around the FC region is steeper along the R_{AB} axis than along the R_{BC} axis. Subsequently, R_{BC} continues decreasing and R_{AB} oscillates around the equilibrium bond length with a frequency of 97 cm^{-1} until R_{BC} eventually reaches the equilibrium bond length at 360 fs. These observations indicate asynchronous bond formation as in path 1, with the covalent bond formed earlier in the Au–Au pair with a shorter distance in the ground state.

We note that the temporal changes in R_{BC} are correlated with the temporal oscillations of R_{AB}. In the time range from 0 fs to 60 fs, both R_{AB} and R_{BC} rapidly decrease by 0.35 Å and 0.16 Å, respectively. In the subsequent time range, from 60 fs to 260 fs, R_{AB} increases by 0.09 Å, whereas R_{BC} continues to decrease, but only by 0.12 Å and with a much lower rate than in the range t < 60 fs. Then, in the time range from 260 fs to 360 fs, both R_{AB} and R_{BC} decrease, by 0.03 Å and 0.25 Å, respectively, such that the rate of decrease of R_{BC} recovers to its initial rate. This correlation between the changes in R_{AB} and R_{BC} indicates that the symmetric stretching mode of the gold trimer complex mediates the bond

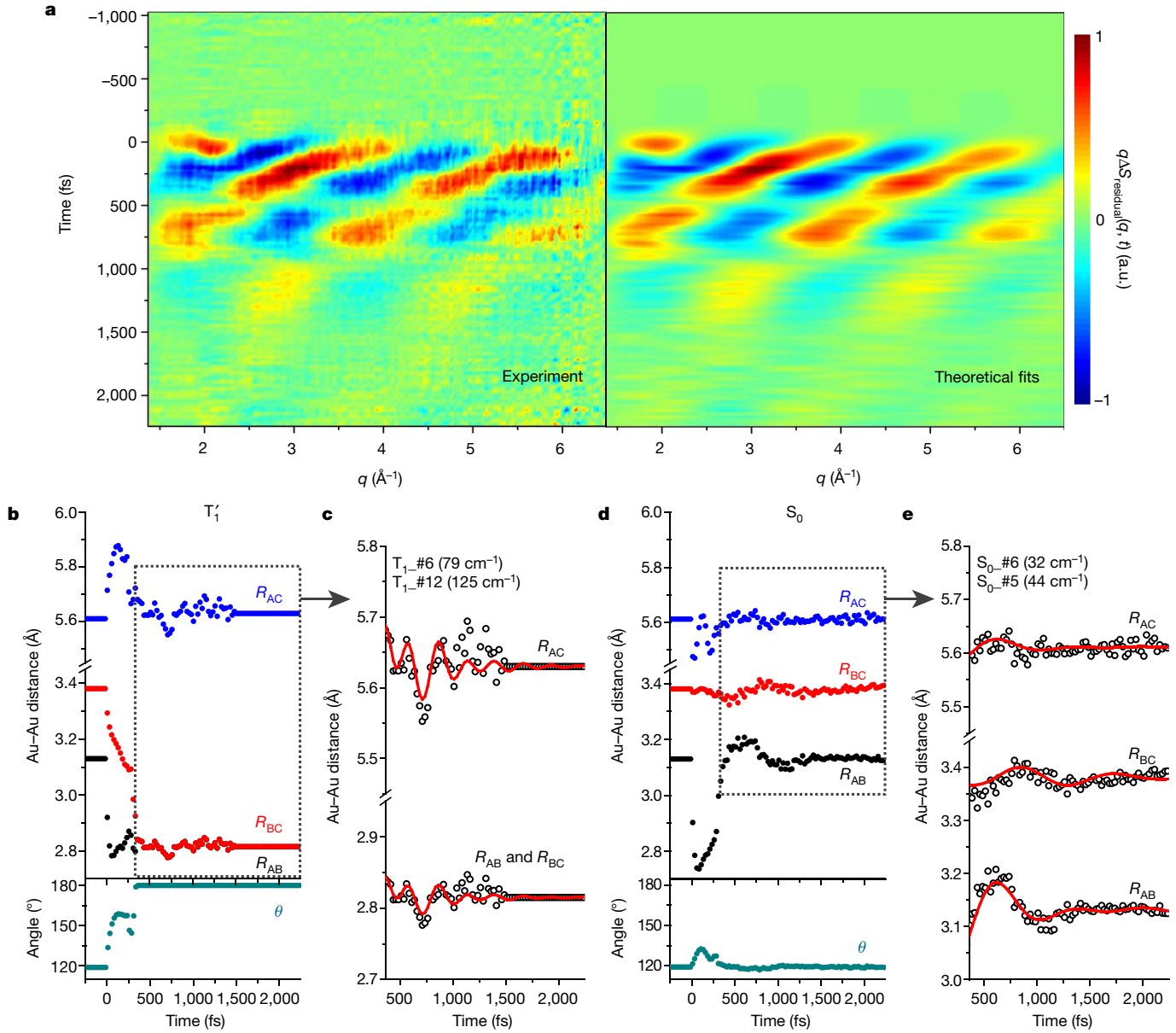


Fig. 2 | Structural analysis using residual difference scattering curves. **a**, Experimental residual difference scattering curves, $q\Delta S_{residual}(q, t)$ (left) and their theoretical fits (right) obtained from the structural analysis performed by considering wavepacket motions in the states S_0 and T_1'. **b**, **d**, Top, time-dependent Au–Au distances $R_{AB}(t)$ (black), $R_{BC}(t)$ (red) and $R_{AC}(t)$ (blue); and bottom, Au–Au–Au angle, (θ, teal) of T_1' (**b**) and S_0 (**d**), determined from the structural analysis. **c**, **e** Magnified views of the structure at $t > 360$ fs for T_1' (**c**) and S_0 (**e**). The measured $R_{AB}(t)$, $R_{BC}(t)$ and $R_{AC}(t)$ (black open circles) are fitted by a sum of two damping cosine functions (red lines), the frequencies of which are given at the top left. As described in the text, specific normal modes of T_1' and S_0 were assigned to these oscillations of the Au–Au distances.

formation. Detailed structural changes of the gold trimer complex associated with the initial wavepacket motion on the PES of T_1' are summarized in Fig. 3b.

The trajectory of the ground-state wavepacket in S_0 is represented in Fig. 3c and its projection onto the R_{AB}–R_{BC} plane is shown in Fig. 3c and Supplementary Fig. 5b. The ground-state wavepacket is generated by two interactions with the pump pulse, that is, via resonant, impulsive stimulated Raman scattering[29], and then within 100 fs is seen to move in the direction of decreasing R_{AB} and increasing θ. This initial motion should reflect the initial structural changes occurring in the excited state—that is, the ultrafast bond formation and the bent-to-linear transformation (see Supplementary Information for details). Detailed structural changes associated with the initial wavepacket motion on the PES of S_0 are shown in Fig. 3d.

After the initial motions of the wavepackets in the ground and excited states as described above, at later times ($t > 360$ fs) the wavepackets oscillate around their equilibrium structures. Molecular vibrations play an important part in the progress of chemical reactions by providing atomic motions along the reaction coordinates and are often discussed as key parameters in the interpretation of reaction dynamics measured with various time-resolved spectroscopies[1–5,19,27,28]. The temporal changes of the structural parameters of T_1' and S_0 after 360 fs are shown in Fig. 2c, e, respectively. It can be seen that all the structural parameters simply oscillate around their own equilibrium values, without any major changes observed in the wavepacket motion at earlier times. To characterize these oscillations at later times, we fitted $R_{AB}(t)$, $R_{BC}(t)$ and $R_{AC}(t)$ of T_1' and S_0 obtained from the structural analysis with various combinations of the vibrational normal modes, that is, the

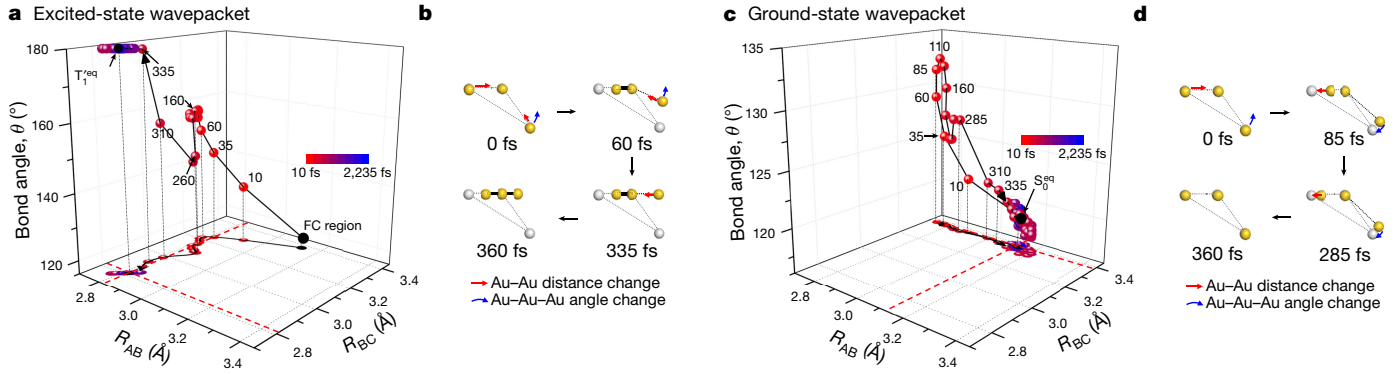


Fig. 3 | Trajectories of the excited-state and ground-state wavepackets determined from TRXL data. a, **c** Motions of the excited-state wavepacket in T_1' (**a**) and the ground-state wavepacket in S_0 (**c**) represented in the multidimensional nuclear coordinates R_{AB} versus R_{BC} versus θ. The projection of the wavepacket motion onto the R_{AB}–R_{BC} plane is shown at the bottom. The equilibrium distances of R_{AB} and R_{BC} in T_1' and S_0 are indicated by the red dashed lines. The positions of the wavepacket at measured time points are indicated by dots, the colours of which represent time delays given by the colour scale. Several representative time delays, given in femtoseconds, are shown next to the corresponding wavepacket position. The black curves connect the dots, ordered by time; they correspond to the trajectory of the wavepacket over time. **b**, **d** Transient structures of T_1' (**b**) and S_0 (**d**) at representative time delays. The Au atoms at each time delay are represented by yellow dots, and the Au atoms in the FC region are represented by grey dots. In **b**, the covalent bonds formed in the excited state are indicated by the black solid lines. The change in interatomic distance and angle are indicated by red and blue arrows, respectively. The ligands are omitted for simplicity. Structural changes are exaggerated for clarity.

symmetric stretching, asymmetric stretching and bending modes. For T_1', a sum of two symmetric stretching modes with frequencies of 79 cm^{-1} and 125 cm^{-1} gives a satisfactory fit to the temporal changes in Au–Au distances, as shown in Fig. 2c. Accordingly, those two oscillations at 79 cm^{-1} and 125 cm^{-1} are assigned to two symmetric stretching modes of T_1': T_1_#6 (theoretical frequency, 63 cm^{-1}) and T_1_#12 (theoretical frequency, 92 cm^{-1}) (see Extended Data Fig. 6), respectively, identified by density functional theory (DFT) calculations (see Methods and Extended Data Fig. 6 for details). Similarly, from the fitting of the temporal changes of the structural parameters of S_0, a symmetric stretching mode with frequency 32 cm^{-1} and an asymmetric stretching mode with frequency 44 cm^{-1} were identified, as shown in Fig. 2e, and assigned to S_0_#6 (theoretical frequency, 58 cm^{-1}) and S_0_#5 (theoretical frequency, 43 cm^{-1}) of S_0, respectively (see Methods and Supplementary Information for details).

The trajectories of the wavepackets in T_1' and S_0 at later times (>360 fs) are shown in the nuclear coordinates R_{AB} versus R_{BC} in Fig. 4a, b, respectively. The displacements of the wavepackets from the equilibrium structures are represented by the sum of structural changes along the two normal coordinates of the activated vibrational modes (T_1_#6 and T_1_#12 for T_1'; S_0_#6 and S_0_#5 for S_0). As shown in Fig. 4a, b, both wavepackets oscillate with respect to the normal coordinates of the activated vibrational modes and eventually approach their equilibrium structures with vibrational dephasing. Therefore, we conclude that in the late time range, both excited-state and ground-state wavepackets exhibit harmonic oscillations around the equilibrium structures.

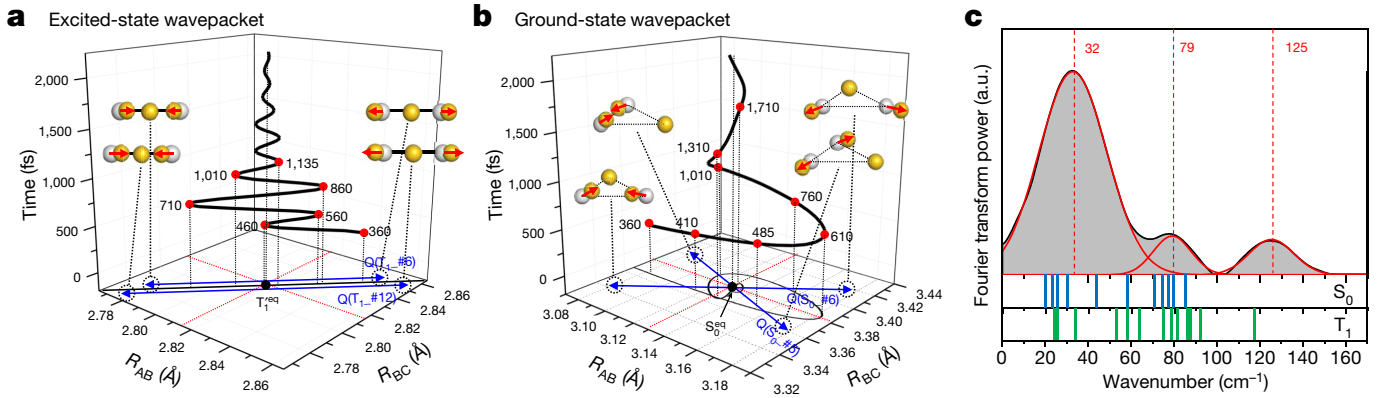


Fig. 4 | Harmonic oscillations of the ground-state and excited-state wavepackets at t > 360 fs. a, **b** Later-time (>360 fs) trajectories of the excited-state wavepacket in T_1' (**a**) and the ground-state wavepacket in S_0 (**b**), represented in the multidimensional nuclear coordinates R_{AB} versus R_{BC} versus time (t). The wavepacket trajectories are indicated by black curves. The wavepacket positions at several representative time delays (given in femtoseconds) are indicated by red dots. The equilibrium distances of R_{AB} and R_{BC} in T_1' and S_0 are indicated by the red dotted lines in **a** and **b**, respectively. The normal coordinates Q of the two symmetric stretching modes for T_1', $Q(T_1\_\#6)$ and $Q(T_1\_\#12)$ (**a**), and the symmetric and asymmetric stretching modes for S_0, $Q(S_0\_\#6)$ and $Q(S_0\_\#5)$ (**b**) are indicated by blue arrows. At the end of each arrow, the representative structure, with Au atoms as yellow spheres, is shown to indicate the displacements of three Au atoms according to the corresponding normal coordinate, and the positions of the Au atoms in the equilibrium structures are represented by grey spheres. The red arrows in the representative structures indicate the displacement vectors of the Au atoms for each mode; they are exaggerated for clarity. In **a**, the normal coordinates—exactly on the diagonal direction of the R_{AB}–R_{BC} plane—are each slightly displaced for clarity. The projections of the trajectories onto the R_{AB}–R_{BC} plane are shown in Supplementary Fig. 5c, d, respectively. **c**, Averaged Fourier power spectrum of $q\Delta S_{residual}(q, t)$ at later times (>360 fs). The peak positions were determined to occur at 32 cm^{-1}, 79 cm^{-1} and 125 cm^{-1}, by fitting the power spectrum with the sum of three Gaussian functions, represented by the red curves. The vertical bars below the Fourier spectrum indicate the DFT-calculated frequencies of the normal modes of S_0 (blue) and T_1 (green) in the frequency range 20–170 cm^{-1}.

Importantly, these harmonic oscillations, manifested by the motions of wavepackets, can be unambiguously assigned to specific normal modes, because the TRXL signal contains information on the vibrational frequency as well as on the atomic-level movement pattern of a vibrational mode (Extended Data Fig. 4).

These assignments can be compared with those made on the basis of only the frequencies of normal modes. The oscillation frequencies can be easily extracted from the Fourier power spectrum of $q\Delta S_{\mathrm{residual}}(q,t)$ at later times. The Fourier transform spectrum shown in Fig. 4c shows a major peak at 32 cm^{-1} and minor peaks at 79 cm^{-1} and 125 cm^{-1}. As detailed in Supplementary Information, if one follows the normal practice of choosing the normal mode with the most similar frequency as the 32 cm^{-1} oscillation from among the 11 modes identified by DFT calculations for S_0 and T_1, a bending mode of T_1 with a frequency of 33 cm^{-1} would be incorrectly chosen. Also, the additional mode (an asymmetric stretching mode with a frequency of 44 cm^{-1}) discovered by TRXL data analysis would not be identified in the Fourier transform spectrum. Therefore, the results of this work showcase that the assignment of the observed oscillations to specific vibrational modes on the basis of a simple numerical comparison of vibrational frequencies is susceptible to misinterpretation for polyatomic molecules, which have many normal modes.

These observations illustrate that femtosecond TRXL can map the real-time trajectories of nuclear wavepackets and thereby identify how vibrational motion drives asynchronous bond formation. Although in this study we have tracked only the motions of highly scattering gold atoms, it is in principle also possible to visualize the motions of lighter atoms such as carbon and nitrogen—especially once next-generation X-ray sources such as LCLS-II HE[30] become available (see the simulation for O_3 presented in Supplementary Information).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2417-3.

1. Krause, J. L., Whitnell, R. M., Wilson, K. R., Yan, Y. J. & Mukamel, S. Optical control of molecular dynamics – molecular cannons, reflectrons, and wave-packet focusers. *J. Chem. Phys.* **99**, 6562–6578 (1993).
2. Kukura, P., McCamant, D. W. & Mathies, R. A. Femtosecond stimulated Raman spectroscopy. *Annu. Rev. Phys. Chem.* **58**, 461–488 (2007).
3. McClure, S. D., Turner, D. B., Arpin, P. C., Mirkovic, T. & Scholes, G. D. Coherent oscillations in the PC577 cryptophyte antenna occur in the excited electronic state. *J. Phys. Chem. B* **118**, 1296–1308 (2014).
4. Cho, S. et al. Coherence in metal–metal-to-ligand-charge-transfer excited states of a dimetallic complex investigated by ultrafast transient absorption anisotropy. *J. Phys. Chem. A* **115**, 3990–3996 (2011).
5. Baumert, T., Engel, V., Rottgermann, C., Strunz, W. T. & Gerber, G. Femtosecond pump–probe study of the spreading and recurrence of a vibrational wavepacket in Na_2. *Chem. Phys. Lett.* **191**, 639–644 (1992).
6. Lin, J. J., Zhou, J. G., Shiu, W. C. & Liu, K. P. State-specific correlation of coincident product pairs in the F + CD_4 reaction. *Science* **300**, 966–969 (2003).
7. Pan, H. L., Wang, F. Y., Czako, G. & Liu, K. P. Direct mapping of the angle-dependent barrier to reaction for Cl + CHD_3 using polarized scattering data. *Nat. Chem.* **9**, 1175–1180 (2017).
8. Ren, Z. F., Sun, Z. G., Zhang, D. H. & Yang, X. M. A review of dynamical resonances in A + BC chemical reactions. *Rep. Prog. Phys.* **80**, 026401 (2017).
9. Kim, K. H. et al. Direct observation of bond formation in solution with femtosecond X-ray scattering. *Nature* **518**, 385–389 (2015).
10. Biasin, E. et al. Femtosecond X-ray scattering study of ultrafast photoinduced structural dynamics in solvated $[Co(terpy)_2]^{2+}$. *Phys. Rev. Lett.* **117**, 013002 (2016).
11. Haldrup, K. et al. Ultrafast X-ray scattering measurements of coherent structural dynamics on the ground-state potential energy surface of a diplatinum molecule. *Phys. Rev. Lett.* **122**, 063001 (2019).
12. van Driel, T. B. et al. Atomistic characterization of the active-site solvation dynamics of a model photocatalyst. *Nat. Commun.* **7**, 13678 (2016).
13. Ishikawa, T. et al. A compact X-ray free-electron laser emitting in the sub-angstrom region. *Nat. Photon.* **6**, 540–544 (2012).
14. Kang, H. S. et al. Hard X-ray free-electron laser with femtosecond-scale timing jitter. *Nat. Photon.* **11**, 708–713 (2017).
15. Cui, G. L., Cao, X. Y., Fang, W. H., Dolg, M. & Thiel, W. Photoinduced gold(I)–gold(I) chemical bonding in dicyanoaurate oligomers. *Angew. Chem. Int. Ed.* **52**, 10281–10285 (2013).
16. Pyykkö, P. Theoretical chemistry of gold. *Angew. Chem. Int. Ed.* **43**, 4412–4456 (2004).
17. Sohn, S. H., Heo, W., Lee, C., Kim, J. & Joo, T. Electronic and structural dynamics of dicyanoaurate trimer in excited state. *J. Phys. Chem. A* **123**, 6904–6910 (2019).
18. Iwamura, M., Nozaki, K., Takeuchi, S. & Tahara, T. Real-time observation of tight Au–Au bond formation and relevant coherent motion upon photoexcitation of $[Au(CN)_2^-]$ oligomers. *J. Am. Chem. Soc.* **135**, 538–541 (2013).
19. Kjær, K. S. et al. Finding intersections between electronic excited state potential energy surfaces with simultaneous ultrafast X-ray scattering and spectroscopy. *Chem. Sci.* **10**, 5749–5760 (2019).
20. Stankus, B. et al. Ultrafast X-ray scattering reveals vibrational coherence following Rydberg excitation. *Nat. Chem.* **11**, 716–721 (2019).
21. Wolf, T. J. A. et al. The photochemical ring-opening of 1,3-cyclohexadiene imaged by ultrafast electron diffraction. *Nat. Chem.* **11**, 504–509 (2019).
22. Glownia, J. M. et al. Self-referenced coherent diffraction X-ray movie of angstrom- and femtosecond-scale atomic motion. *Phys. Rev. Lett.* **117**, 153003 (2016).
23. Yang, J. et al. Diffractive imaging of coherent nuclear motion in isolated molecules. *Phys. Rev. Lett.* **117**, 153002 (2016).
24. Yang, J. et al. Imaging CF_3I conical intersection and photodissociation dynamics with ultrafast electron diffraction. *Science* **361**, 64–67 (2018).
25. Debnarova, A., Techert, S. & Schmatz, S. Ab initio treatment of time-resolved X-ray scattering: application to the photoisomerization of stilbene. *J. Chem. Phys.* **125**, 224101 (2006).
26. Henriksen, N. E. & Moller, K. B. On the theory of time-resolved X-ray diffraction. *J. Phys. Chem. B* **112**, 558–567 (2008).
27. Katayama, T. et al. Tracking multiple components of a nuclear wavepacket in photoexcited Cu(I)-phenanthroline complex using ultrafast X-ray spectroscopy. *Nat. Commun.* **10**, 3606 (2019).
28. Lemke, H. T. et al. Coherent structural trapping through wavepacket dispersion during photoinduced spin state switching. *Nat. Commun.* **8**, 15342 (2017).
29. Pollard, W. T. & Mathies, R. A. Analysis of femtosecond dynamic absorption spectra of nonstationary states. *Annu. Rev. Phys. Chem.* **43**, 497–523 (1992).
30. Schoenlein, R. W., Boutet, S., Minitti, M. P. & Dunne, A. M. The Linac Coherent Light Source: recent developments and future plans. *Appl. Sci.* **7**, 850 (2017).

Methods

TRXL experiments at PAL-XFEL

TRXL experiments were performed at the XSS beamline of PAL-XFEL (the Pohang Accelerator Laboratory X-ray free-electron laser). In the TRXL experiment, the sample solution of $[Au(CN)_2^-]_3$ was excited by an optical laser pulse to initiate a photoinduced reaction of the sample molecules and a time-delayed X-ray pulse was used to probe the progress of the reaction. Femtosecond laser pulses at the centre wavelength of 800 nm were generated from a Ti:sapphire regenerative amplifier and converted to 100-fs pulses at a wavelength of 267 nm by third-harmonic generation. The laser beam was focused by a lens to a spot of 200-μm diameter at the sample position, yielding a laser fluence of 1.5 mJ mm^{-2}. Femtosecond X-ray pulses were generated from an X-ray free-electron laser (XFEL) by self-amplified spontaneous emission. The X-ray pulses have a centre energy of 12.7 keV with a narrow energy bandwidth ($\Delta E/E = 0.3\%$). The X-ray beam was focused to a spot 40 μm in diameter at the sample position. The laser and X-ray beams were overlapped at the sample position with a crossing angle of 10°. The X-ray scattering patterns from the photoexcited $[Au(CN)_2^-]_3$ solution generated by the X-ray pulses were measured with an area detector (MX225-HS, Rayonix) over a q range of 1.37 Å^{-1} to 6.5 Å^{-1} with a sample-to-detector distance of 46 mm.

The TRXL data were measured at various time delays in the range −1,040 fs to 2,235 fs with a time step of 25 fs, yielding a total of 132 time delays. The laser-off images were acquired with the X-ray pulse arriving 20 ps earlier than the laser pulse (that is, with a −20 ps time delay) to probe the (unexcited) molecules in the ground state while ensuring the same average temperature of the sample solution. These laser-off images were repeatedly measured before every laser-on image and were subtracted from the laser-on images to yield time-resolved difference scattering patterns of the $[Au(CN)_2^-]_3$ solution. Each scattering image was obtained with a single X-ray pulse and, to achieve a signal-to-noise ratio sufficient for data analysis, around 2,800 images were acquired at each time delay. The resultant time-resolved difference scattering curves are shown in Extended Data Fig. 2a.

For the sample, we used an aqueous solution of a gold oligomer complex, $[Au(CN)_2^-]_n$. In the solution of $Au(CN)_2^-$ at the 300 mM concentration used in this work, the $[Au(CN)_2^-]_3$ trimers are dominantly present compared with dimers or monomers of $Au(CN)_2^-$. The sample solution was excited by the laser pulses of 267-nm wavelength. The sample solution was circulated through a nozzle with a 100-μm-thick aperture. To supply a fresh sample for every laser and X-ray shot, the flow velocity of the sample was set to be over 3 m s^{-1}. To prevent the scattering signal from contamination by radiation-damaged sample molecules, the sample in the reservoir was replaced with a fresh one whenever the transient signal measured at 100 ps was no longer reproduced. Even if the transient signal at 100 ps did not change, the sample in the reservoir was regularly replaced (every 2–3 h of measurement) to ensure the supply of fresh samples.

TRXL experiments at SACLA

The TRXL experiments were also performed at the BL3 beamline of SACLA (the SPring-8 ångstrom compact free-electron laser). In the TRXL experiment, the photoinduced reaction of the gold complex was initiated by an optical laser pulse and its progress was probed by a time-delayed X-ray pulse. Femtosecond laser pulses at a centre wavelength of 800 nm were generated from a Ti:sapphire regenerative amplifier and converted to 200-fs pulses at a wavelength of 267 nm by third-harmonic generation. The laser beam was focused by a lens to a spot of 300-μm diameter at the sample position, yielding a laser fluence of about 2 mJ mm^{-2}. Femtosecond X-ray pulses were generated from an XFEL by self-amplified spontaneous emission. The X-ray pulses have a centre energy of 15 keV with a narrow energy bandwidth ($\Delta E/E = 0.6\%$). The X-ray beam was focused to a spot of diameter 200 μm at the sample position. The laser and X-ray beams were overlapped at the sample position with a crossing angle of 10°. The X-ray scattering patterns from the photoexcited $[Au(CN)_2^-]_3$ solution generated by the X-ray pulses were measured with an area detector (LX255-HS, Rayonix) over a q range of 1.37 Å^{-1} to 6.5 Å^{-1} with a sample-to-detector distance of 30 mm. To improve the time resolution of the TRXL measurements, a timing monitor installed at SACLA was used. The TRXL data were measured at various time delays from −740 fs to 2,260 fs with a time step of 25 fs, yielding a total of 121 time delays. At each time delay, about 2,000 images were accumulated. The same data-acquisition scheme as that used at PAL-XFEL was used for the TRXL experiment at SACLA. The resultant time-resolved difference scattering curves are shown in Extended Data Fig. 2b.

As can be seen in Extended Data Fig. 2a, b, the two TRXL datasets measured at PAL-XFEL and SACLA are nearly identical, except for the time resolution (170 fs at PAL-XFEL and 320 fs at SACLA), indicating that the difference scattering signals are highly reproducible at either facility. In this work, we primarily used the TRXL dataset measured at PAL-XFEL, which has better time resolution and signal-to-noise ratio.

To eliminate the contribution of solvent heating, the difference scattering signal of 40 mM $FeCl_3$ solution was measured from a separate TRXL experiment, as shown in Extended Data Fig. 8, with the same experimental conditions used in the TRXL experiment on the gold trimer complex at SACLA (See Supplementary Information for details).

Singular value decomposition

To extract the kinetics from the measured TRXL data of $[Au(CN)_2^-]_3$, we applied an SVD analysis. To do so, we built an $n_q \times n_t$ data matrix, A, the column vectors of which are experimental time-resolved difference scattering curves, where n_q is the number of q points in the difference scattering curves and n_t is the number of time-delay points. By SVD, the matrix A is decomposed into three matrices satisfying the relationship $A = USV^T$ (where V^T is the transpose of matrix V). U is an $n_q \times n_t$ matrix the column vectors of which are called the left singular vectors (LSVs) of A, V is an $n_t \times n_t$ matrix the column vectors of which are the RSVs of A, and S is a diagonal $n_t \times n_t$ matrix the diagonal elements of which are called the singular values of A. The matrices U and V follow the relationships $U^TU = I_{n_t}$ and $V^TV = I_{n_t}$, respectively, where I_{n_t} is an $n_t \times n_t$ identity matrix. The LSVs represent time-independent q spectra, the RSVs represent the time-dependent amplitude changes of the corresponding LSVs, and the singular values represent the weights of the corresponding LSVs and RSVs. The singular values are ordered such that $s_1 \geq s_2 \geq \ldots \geq s_n \geq 0$, and so (both left and right) singular vectors in the left-hand columns have larger contributions to the experimental data matrix A. The first and second RSVs shown in Extended Data Fig. 2c were well fitted with the convolution of a Gaussian function with a full width at half-maximum (FWHM) of 170 ± 50 fs and an exponential function with a time constant of 1.1 ± 0.1 ps. The errors are the standard errors of the mean determined from around 2,800 independent measurements.

Residual difference scattering curves, $q\Delta S_{\mathrm{residual}}(q, t)$

To more effectively visualize the scattering intensities arising from wavepacket motions, we extracted residual difference scattering curves, $q\Delta S_{\mathrm{residual}}(q, t)$, from raw experimental difference scattering curves, $q\Delta S(q, t)$, of photoexcited $[Au(CN)_2^-]_3$. $\Delta S(q, t)$ can be represented as follows:

$$\begin{aligned}\Delta S(q,t) = {} & c_{T_1'}(t)[S_{T_1'(t)}(q) - S_{S_0^{eq}}(q)] + c_{T_1}(t)[S_{T_1(t)}(q) - S_{S_0^{eq}}(q)] \\ & + [c_{T_1'}(t) + c_{T_1}(t)][S_{S_0(t)}(q) - S_{S_0^{eq}}(q)] + \Delta S_{\mathrm{heat}}(q,t),\end{aligned} \quad (1)$$

where $S_{S_0(t)}(q)$, $S_{T_1'(t)}(q)$ and $S_{T_1(t)}(q)$ are the scattering patterns of the instantaneous structures of the S_0, T_1' and T_1 states, respectively, that evolve following vibrational wavepacket motions. $S_{S_0^{eq}}(q)$ is the

scattering pattern of the equilibrium S_0 structure, and $c_{T_1'}(t)$ and $c_{T_1}(t)$ are the time-dependent relative populations of the T_1' and T_1 states, respectively. $\Delta S_{heat}(q, t)$ represents the change in scattering intensity induced by solvent heating.

Alternatively, $\Delta S(q, t)$ can be represented as the sum of scattering contributions of: (1) the dynamics of the T_1'-to-T_1 transition, $\Delta S_{transit}(q, t)$; (2) the temporal oscillations of scattering intensities owing to vibrational wavepacket motions in the S_0, T_1' and T_1 states—that is, the residual difference scattering curves, $\Delta S_{residual}(q, t)$; and (3) the solvent heating, $\Delta S_{heat}(q,t)$, as

$$\Delta S(q, t) = \Delta S_{transit}(q, t) + \Delta S_{residual}(q, t) + \Delta S_{heat}(q, t), \quad (2)$$

where

$$\Delta S_{transit}(q, t) = c_{T_1'}(t)[S_{T_1'^{eq}}(q) - S_{S_0^{eq}}(q)] + c_{T_1}(t)[S_{T_1^{eq}}(q) - S_{S_0^{eq}}(q)], \quad (3)$$

$$\begin{aligned}\Delta S_{residual}(q, t) = {} & [c_{T_1'}(t) + c_{T_1}(t)][S_{S_0(t)}(q) - S_{S_0^{eq}}(q)] \\ & + c_{T_1'}(t)[S_{T_1'(t)}(q) - S_{T_1'^{eq}}(q)] \\ & + c_{T_1}(t)[S_{T_1(t)}(q) - S_{T_1^{eq}}(q)]\end{aligned} \quad (4)$$

and $S_{T_1'^{eq}}(t)$ and $S_{T_1^{eq}}(q)$ are the scattering intensities arising from the equilibrium structure of the T_1' and T_1 states, respectively. To extract the residual difference scattering curves, $q\Delta S_{residual}(q, t)$, we subtracted the contributions of: (1) the T_1'-to-T_1 transition dynamics; and (2) the solvent heating from $\Delta S(q, t)$, as described in the following.

We note that $\Delta S_{transit}(q, t)$ shows the dynamics described by an exponential with a time constant of 1.1 ps, whereas $\Delta S_{residual}(q, t)$ exhibits temporal oscillation owing to wavepacket motions. To extract $\Delta S_{residual}(q, t)$ from $\Delta S(q, t)$, we examined the RSVs obtained from the SVD analysis described in the previous section. As can be seen in Extended Data Fig. 2d, the first two RSVs exhibit exponential dynamics with superimposed temporal oscillations, whereas other RSVs oscillate only around zero. By removing the exponential components from the first two RSVs, we can remove the scattering contribution of the T_1'-to-T_1 dynamics, $\Delta S_{transit}(q, t)$, and the contribution of the solvent heating, $\Delta S_{heat}(q, t)$. In fact, as can be seen in Extended Data Fig. 7, the TRXL data from our previous TRXL study[9] on $[Au(CN)_2^-]_3$, which involves only the contributions from $\Delta S_{transit}(q, t)$ and $\Delta S_{heat}(q, t)$, can be well explained by the first two RSVs. Therefore, the removal of the exponential components from the first two RSVs of $\Delta S(q, t)$ removes $\Delta S_{transit}(q, t)$ and $\Delta S_{heat}(q, t)$, leaving only $\Delta S_{residual}(q, t)$.

To eliminate the exponential components from the first two singular vectors, we defined new matrices, U', V' and S', which contain only the first two column vectors of U, V and S, respectively. In other words, U' is an $n_q \times 2$ matrix containing only the first two LSVs of U, S' is a 2×2 diagonal matrix containing only the first two singular values of S, and V' is an $n_t \times 2$ matrix containing only the first two RSVs of V. We then defined a matrix C that represents the exponential temporal profiles of the two RSVs. Elements of the matrix C were calculated as follows:

$$\mathbf{c}_1(t) = \mathrm{IRF}(t) \otimes [\exp(-t/1.1\mathrm{ps})\theta(t)] \text{ and}$$

$$\mathbf{c}_2(t) = \mathrm{IRF}(t) \otimes [(1 - \exp(-t/1.1\mathrm{ps}))\theta(t)], \quad (5)$$

where $\mathbf{c}_1(t)$ and $\mathbf{c}_2(t)$ are the first and second column vectors of C, IRF(t) is the instrument response function determined from the fitting of the first and second RSVs shown in Extended Data Fig. 2c, $\theta(t)$ is the Heaviside step function and $\otimes$ is the convolution operator. Then, a 2×2 parameter matrix P is defined to relate C to V'. Elements of P were adjusted to minimize the discrepancy between V' and CP. As a result, the exponential components contained in the first two RSVs (V') can be represented by an optimized CP. Then, time-dependent scattering intensities, which are governed by the exponential dynamics of the first two RSVs, were calculated by the following relationship, $A' = U'S'(CP)^T$, where the scattering intensities are column vectors of the A' matrix. Finally, A' was subtracted from A, giving $\Delta S_{residual}(q, t)$ as column vectors of the matrix $A - A'$.

Structural analysis using residual difference scattering curves

As shown in Fig. 2, we fitted $\Delta S_{residual}(q, t)$ by the theoretical difference scattering curves, $\Delta S_{theory}(q, t)$, to extract the temporal changes of the individual interatomic Au–Au distances—$R_{AB}(t)$, $R_{BC}(t)$ and $R_{AC}(t)$—from the experimental residual difference scattering curves, $\Delta S_{residual}(q, t)$. To do so, we constructed theoretical difference scattering curves, $\Delta S_{theory}(q, t)$, as follows:

$$\begin{aligned}\Delta S_{theory}(q, t) = {} & [c_{T_1'}(t) + c_{T_1}(t)]\Delta S_{S_0}(q, t) + c_{T_1'}(t)\Delta S_{T_1'}(q, t) \\ & + c_{T_1}(t)\Delta S_{T_1}(q, t).\end{aligned} \quad (6)$$

In equation (6), $\Delta S_X(q, t)$ is the difference in scattering intensity arising from a transient structure of state X ($S_X(q, t)$; $X = \{S_0, T_1', T_1\}$), and that arising from the equilibrium structure of state X, $S_{X^{eq}}(q)$, calculated by the following equation:

$$\Delta S_X(q, t) = S_X(q, t) - S_{X^{eq}}(q). \quad (7)$$

Scattering intensities arising from the molecular structures of S_0, T_1' and T_1 were calculated using the Debye equation as follows:

$$S(q) = 3F_{Au}^2(q) + 2F_{Au}^2(q)\left(\frac{\sin(qR_{AB})}{qR_{AB}} + \frac{\sin(qR_{BC})}{qR_{BC}} + \frac{\sin(qR_{AC})}{qR_{AC}}\right), \quad (8)$$

where $F_{Au}(q)$ is the atomic form factor of an Au atom. Debye–Waller factors (DWFs) were introduced to consider distributions of interatomic distances: for S_0, arising from the weak Au–Au bonding in S_0, and for T_1', arising from the spatial broadening of the initially created wavepacket, (which is induced by a finite pulse duration of the pump pulse), on the PES of T_1'. DWFs for T_1 were not used, as their use did not improve the fit quality. When including the DWF, here equation (8) is modified to become equation (5) of the Supplementary Information. The DWFs for S_0 and T_1' used in the fitting analysis are shown in Supplementary Table 1.

For the fitting, the discrepancy between $\Delta S_{residual}(q, t)$, and $\Delta S_{theory}(q, t)$ was minimized by independently adjusting the structural parameters (R_{AB}, R_{BC} and θ) of the S_0, T_1' or T_1 states at each time delay, and time-dependent molecular structures were obtained from the fit over the entire time range. At each time delay, the molecular structure was optimized using a maximum likelihood estimation with the χ^2 estimator, which is given by the following equation:

$$\chi^2 = \frac{1}{n_q - p - 1}\sum_i^{n_q}\frac{(c_s\Delta S_{theory}(q_i, t) - \Delta S_{residual}(q_i, t))^2}{\sigma(q_i, t)^2}. \quad (9)$$

Here, n_q is the number of fitted q points, p is the number of fitting parameters, σ is the standard deviation of the data and c_s is the scaling factor between the theoretical and experimental difference curves. The fitting was performed with the MINUIT software package and the error values were obtained with the *MINOS* algorithm in MINUIT.

The TRXL signal can be sensitive to wavepacket motions on any of the structurally distinct S_0, T_1' and T_1 states, and so we examined which state is associated with the residual difference scattering curves. The first, second and third terms in equation (6) correspond to wavepacket motions in S_0, T_1' and T_1, respectively. Depending on the number of participating states, those terms were selectively used. For example, when we considered the wavepacket motion in a single electronic state, we considered only the term corresponding to that electronic state among the three terms in equation (6); the other two

terms were neglected. When we considered wavepacket motions in two electronic states, we considered the corresponding two terms in equation (6) and neglected the remaining term. We first tried structural analysis considering only one of S_0, T_1' and T_1. As shown in Extended Data Fig. 3, the structural analysis performed using only S_0 yielded the best fit to the experimental data among the three cases, but there still remained a discrepancy between the experimental data and the theoretical fits. We therefore considered additional contributions: the best fit to $\Delta S_{\mathrm{residual}}(q, t)$—shown in Fig. 2a and Extended Data Fig. 3—was obtained when T_1' was considered together with S_0, indicating that the residual difference scattering curves arise from wavepacket motions on the PESs of both S_0 and T_1'.

We note that the structural analysis described above was conducted for time delays later than the experimental IRF (>170 fs). The structural analysis for time delays earlier than 170 fs is described in the section 'Structural analysis for transient structures around $t = 0$'. Although we used all of the structural parameters (R_{AB}, R_{BC} and θ) of S_0 and T_1' as fitting parameters in the structural analysis, we observed relatively small structural changes at later times ($t > 360$ fs), compared with structural changes at earlier times; in particular, the changes in R_{AB} and R_{BC} of T_1' showed a strong correlation. Accordingly, we checked whether the experimental data could be satisfactorily fitted even when fewer structural parameters of T_1' were used for fitting in the later time range. To do so, we classified vibrations of T_1' at later times into three types of vibrational motions (symmetric stretching, asymmetric stretching and bending), and the fitting parameters of T_1' were collectively adjusted to simulate each type of motion. For simulating symmetric stretching, the transient structures of T_1' at later times were set to be $R_{\mathrm{AB}} = R_{\mathrm{BC}} = R$ and $\theta = 180°$, leaving R as the only fitting parameter at each time delay. For simulating asymmetric stretching, the parameters of T_1' were adjusted as $R_{\mathrm{AB}} = 2.82\ \text{Å} + R$, $R_{\mathrm{BC}} = 2.82\ \text{Å} - R$ and $\theta = 180°$, also leaving R as the only fitting parameter at each time delay. Finally, for simulating bending, R_{AB} and R_{BC} of T_1' were fixed to 2.82 Å and θ was used as the only fitting parameter at each time delay. As a result, only one fitting parameter of T_1' was used for all the three types of vibrational motions in the late time range ($t > 360$ fs). By contrast, for S_0 we used all three structural parameters, R_{AB}, R_{BC} and θ.

Following this approach, we performed the structural analysis for each of the three types of vibrational motions. As can be seen in Extended Data Fig. 9a–c, the structural analysis considering the symmetric stretching yielded the best fits to the experimental data among the three cases. Also, the fits using symmetric stretching are equally good as the fits obtained by using all the three structural parameters of T_1', as shown in Extended Data Fig. 9d. This result indicates that only symmetric stretching is observed for T_1' in the current TRXL data. We also note that, after 1,500 fs, the transient structures of T_1' were fixed to the values identical to the equilibrium structure of T_1' ($R_{\mathrm{AB}} = 2.82$ Å, $R_{\mathrm{BC}} = 2.82$ Å, $\theta = 180°$) for simplicity, as shown in Fig. 2b, c, because the quality of the fit to $q\Delta S_{\mathrm{residual}}(q, t)$ did not deteriorate even when the structure values of T_1' were set to the equilibrium structure values for all time delays after 1,500 fs.

From the structural analysis described above, we obtained transient structures of S_0 and T_1' in the time range from 170 fs to 2,235 fs, as shown in Fig. 2b–e. To identify the reaction mechanism of ultrafast bond formation in the gold trimer complex, we tracked the wavepacket motion in the excited state by inspecting the structural changes of T_1' at earlier times. For example, we checked the transient structure of T_1' at 185 fs ($R_{\mathrm{AB}} = 2.81$ Å, $R_{\mathrm{BC}} = 3.13$ Å, $\theta = 158°$) and 210 fs ($R_{\mathrm{AB}} = 2.83$ Å, $R_{\mathrm{BC}} = 3.11$ Å, $\theta = 157°$) and found that for these time delays, the values of R_{AB} for T_1' are similar to the equilibrium value, that is, 2.82 Å. By contrast, the values of R_{BC} for T_1' at 185 fs and 210 fs are longer than R_{BC} at equilibrium (2.82 Å). This observation indicates that the covalent bond in the shorter Au–Au pair (of the ground state) is formed sooner, during the earlier time range, whereas the other covalent bond, in the longer Au–Au pair, forms later, supporting path 1 of Fig. 1.

In the structural analysis presented in this work, we used an asymmetric equilibrium structure for S_0. In our previous TRXL study[9] on $[Au(CN)_2^-]_3$—performed with lower time resolution than in this work—the TRXL data were equally well fitted by a symmetric S_0 structure when an appropriate Debye–Waller factor was used. Therefore, in the present work, we also considered the possibility that S_0 has a symmetric structure where R_{AB} is equal to R_{BC}. As shown in Extended Data Fig. 10, the structural analysis using the symmetric equilibrium S_0 structure does not give satisfactory agreement with the experimental data. Thus, by performing the TRXL measurement with higher time resolution and resolving the signatures of molecular vibrations, we confirm that the equilibrium structure of S_0 is asymmetric.

Structural analysis for transient structures around $t = 0$

To extract the wavepacket trajectory and obtain the transient structures of S_0 and T_1' on timescales shorter than the temporal width of the experimental IRF (<170 fs), and to visualize the progress of the bond formation process more clearly and obtain accurate timescales of the earlier bond formation, we performed a structural analysis for transient structures around $t = 0$ considering the convolution of the molecular response with the IRF, instead of conducting the structural analysis directly on the residual difference scattering curves described in the previous section. To do so, we modelled the interatomic Au–Au distances of S_0 and T_1', $R_{\mathrm{AB}}(t)$, $R_{\mathrm{BC}}(t)$ and $R_{\mathrm{AC}}(t)$, from 0–170 fs using the polynomial functions

$$R(t) = \sum_{i=0}^{N} a_{N-i}\, t^{N-i}, \tag{10}$$

where N represents the Nth order polynomial function and a_{N-i} is the coefficient of the polynomial function. Also, constraints were applied to the Au–Au distances calculated by the polynomial functions to smoothly connect the structure at 0 fs (S_0^{eq} = the FC region) and the structure at 185 fs obtained from the structural analysis.

Using the Au–Au distances calculated by the polynomial functions at various time delays, we calculated the theoretical difference scattering intensities, $\Delta S_{\mathrm{theory}}(q, t)$, following equations (6)–(8), then convoluted them with the IRF determined in Extended Data Fig. 2c, IRF(t), using equation (11):

$$\Delta S_{\mathrm{conv}}(q, t) = \Delta S_{\mathrm{theory}}(q, t) \otimes \mathrm{IRF}(t). \tag{11}$$

The resultant convoluted curves, $\Delta S_{\mathrm{conv}}(q, t)$, were compared with the experimental residual difference scattering curves, $\Delta S_{\mathrm{residual}}(q, t)$, in a time range from −140 fs to 160 fs, to optimize the coefficients of the polynomial functions using the χ^2 estimator, given by

$$\chi^2 = \frac{1}{n_q n_t - p - 1} \sum_{i}^{n_q} \sum_{j}^{n_t} \frac{(c_s \Delta S_{\mathrm{conv}}(q_i, t_j) - \Delta S_{\mathrm{residual}}(q_i, t_j))^2}{\sigma_{ij}^2}, \tag{12}$$

where n_q is the number of fitted q points, n_t is the number of fitted time delays ($-140\ \text{fs} \le t \le 160$ fs), p is the number of fitting parameters, σ_{ij} is the standard deviation, and c_s is the scaling factor between the theoretical and experimental difference curves.

The structures around $t = 0$ shown in Fig. 2b, d were obtained using fourth-order polynomial functions because the third-order and fifth-order polynomial functions yielded fitting qualities poorer than and similar to the fourth-order polynomial functions, respectively.

Normal-mode calculation

Geometry optimization and normal-mode calculations were performed using DFT for the S_0 and T_1 states of $[Au(CN)_2^-]_3$, and using time-dependent DFT for the S_1 state of $[Au(CN)_2^-]_3$. For the S_0 state, the PBE0 exchange-correlation functional with empirical dispersion

(Grimme's D3 method) was used to reasonably describe the weak interaction between Au atoms. The ωB97XD functional was used for the S_1 and T_1 states. The aug-cc-pVTZ-PP relativistic effective core potentials were used for Au atoms and the aug-cc-pVDZ basis set was used for C and N atoms. The solvent (water) effect was modelled using the integral equation formalism version of the polarizable continuum model. Normal modes of the T_1 state were calculated for the equilibrium structure that was optimized by the DFT calculation, and the optimized structure was in agreement with the equilibrium structure of T_1 determined from the experimental data. By contrast, from the DFT calculation the optimized structure of S_0 was determined to be symmetric, and therefore, instead, the normal modes of the S_0 state were calculated by a single-point calculation using the asymmetric equilibrium structure refined with the experimental data. Several normal modes for S_0 and T_1 are shown in Extended Data Figs. 5 and 6, respectively. All the calculations were performed using the Gaussian09 program.

Fourier power spectrum of $q\Delta S_{\mathrm{residual}}(q, t)$

The Fourier power spectrum at each q value was obtained by Fourier transform of $q\Delta S_{\mathrm{residual}}(q, t)$ at later times (>360 fs), and each spectrum was normalized at its maximum value. All the normalized spectra at various q values were averaged to yield the averaged Fourier power spectrum shown in Fig. 4c.

Data availability

The datasets generated and analysed here are available from the corresponding author on reasonable request.

Code availability

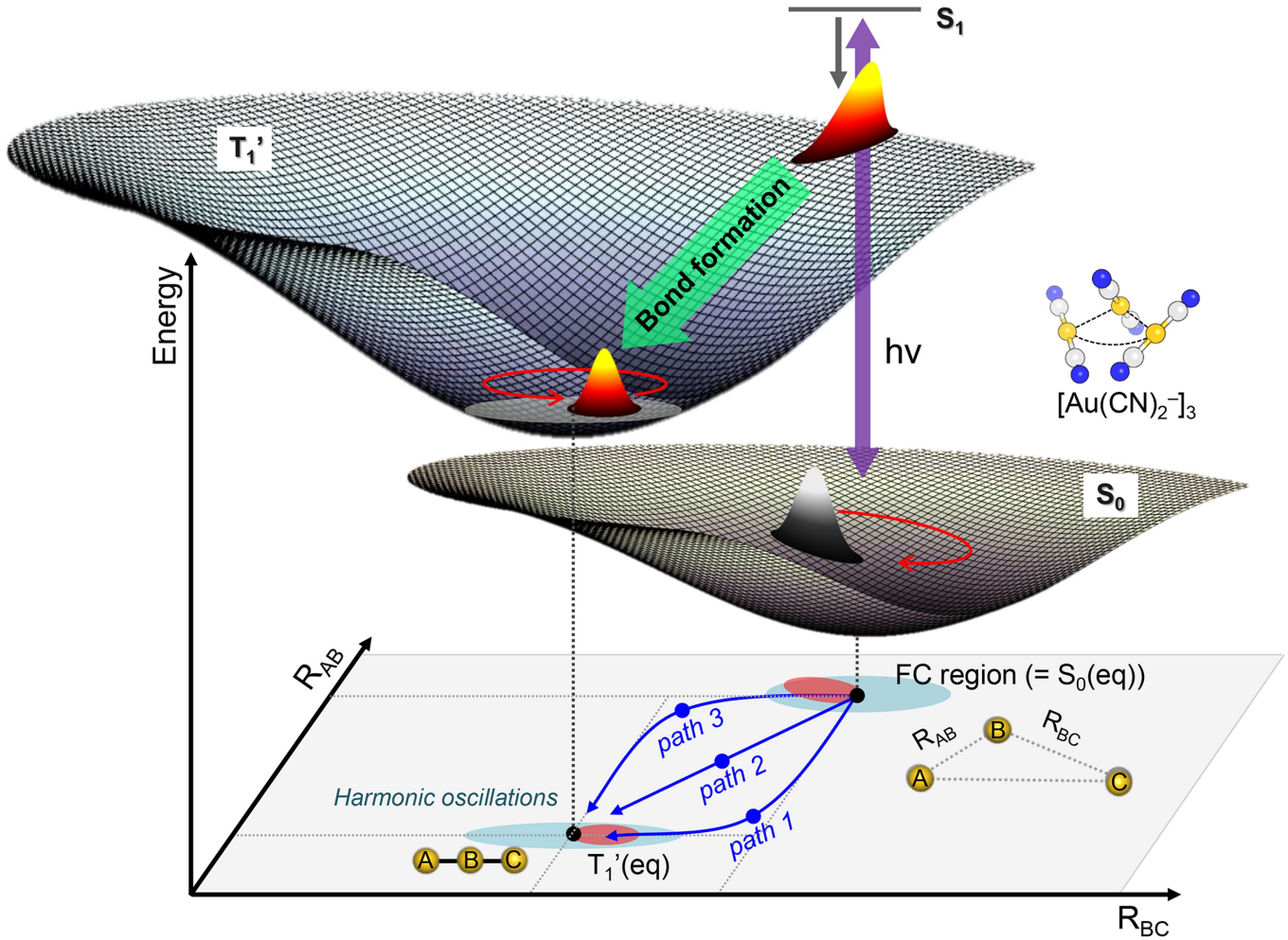
The codes used for the analysis here are available from the corresponding author on reasonable request.

Acknowledgements This work was supported by the Institute for Basic Science (IBS-R004). This work was supported by the X-ray Free-Electron Laser Priority Strategic Program and the Photon and Quantum Basic Research Coordinated Development Program of MEXT, Japan. This work was supported by JSPS KAKENHI grant numbers JP17H06141, JP17H06372, JP17H06438 and JP19H05782. This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2016R1E1A1A01941978). Experiments were performed at the XSS of PAL-XFEL (proposal numbers 2017-2nd-XSS-001 and 2018-2nd-XSS-005), and at the BL3 of SACLA with the approval of the Japan Synchrotron Radiation Research Institute (proposal numbers 2016A8035, 2016A8055, 2016B8056, 2016B8073, 2017A8043, 2017A8053, 2017B8029, 2018A8006, 2018B8015, 2019A8012 and 2019B8025).

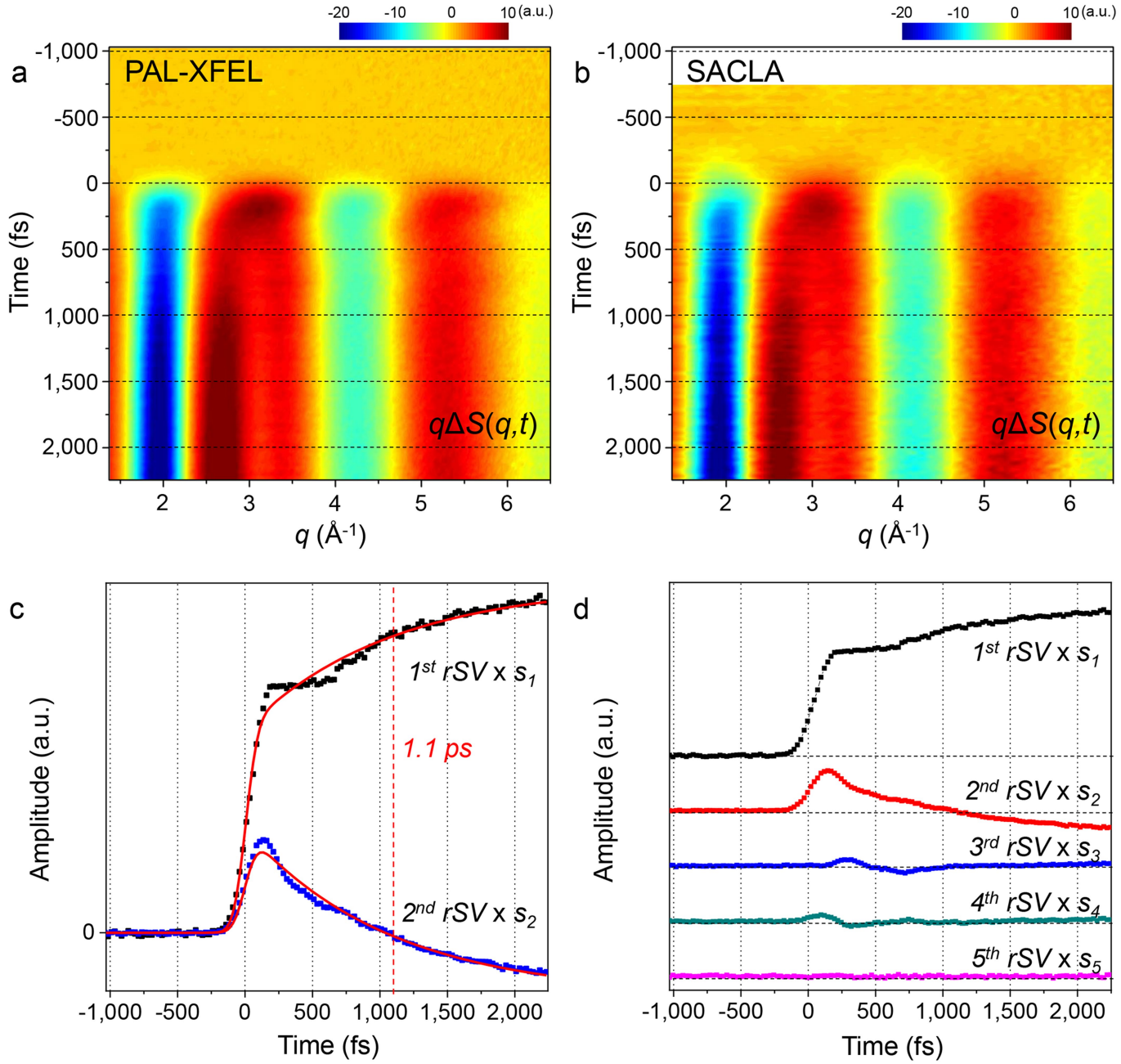
Author contributions H.I. supervised the project; S.-i.A. and H.I. designed the experiment; J.G.K. and H.I. developed the data analysis strategy; J.G.K., S.N., H. Kim, E.H.C., T.S., T.W.K., K.H.K., H. Ki, Jungmin Kim, M.C., Y.L., J.H., K.Y.O., K.I., R.F., J.H.L., J.P., I.E., S.H.C., S.K., M.K., T.K., T.T., S.O., M.Y., S.J.L., S.L., C.W.A., S.C., Jeongho Kim, S.-i.A. and H.I. performed the experiments; J.G.K., H. Kim, E.H.C., K.H.K., D.-S.A. and T.J. analysed the data; J.M. and Joonghan Kim performed quantum chemical calculations; J.G.K., S.N., Jeongho Kim, S.-i.A. and H.I. wrote the manuscript with contributions from all authors.

Competing interests The authors declare no competing interests.

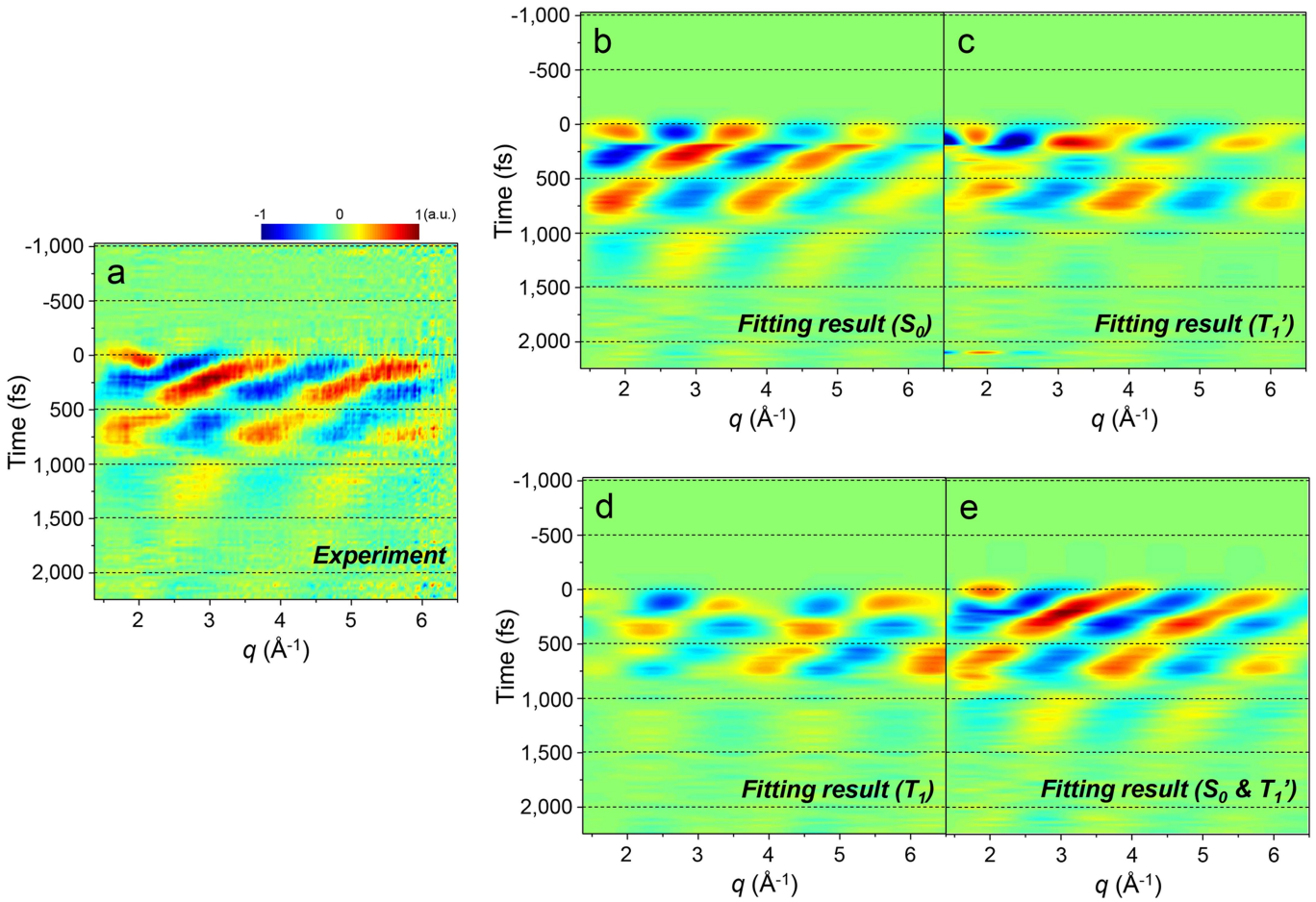
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2417-3.
Correspondence and requests for materials should be addressed to H.I.
Peer review information *Nature* thanks Richard A. Mathies, Martin Meedom Nielsen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



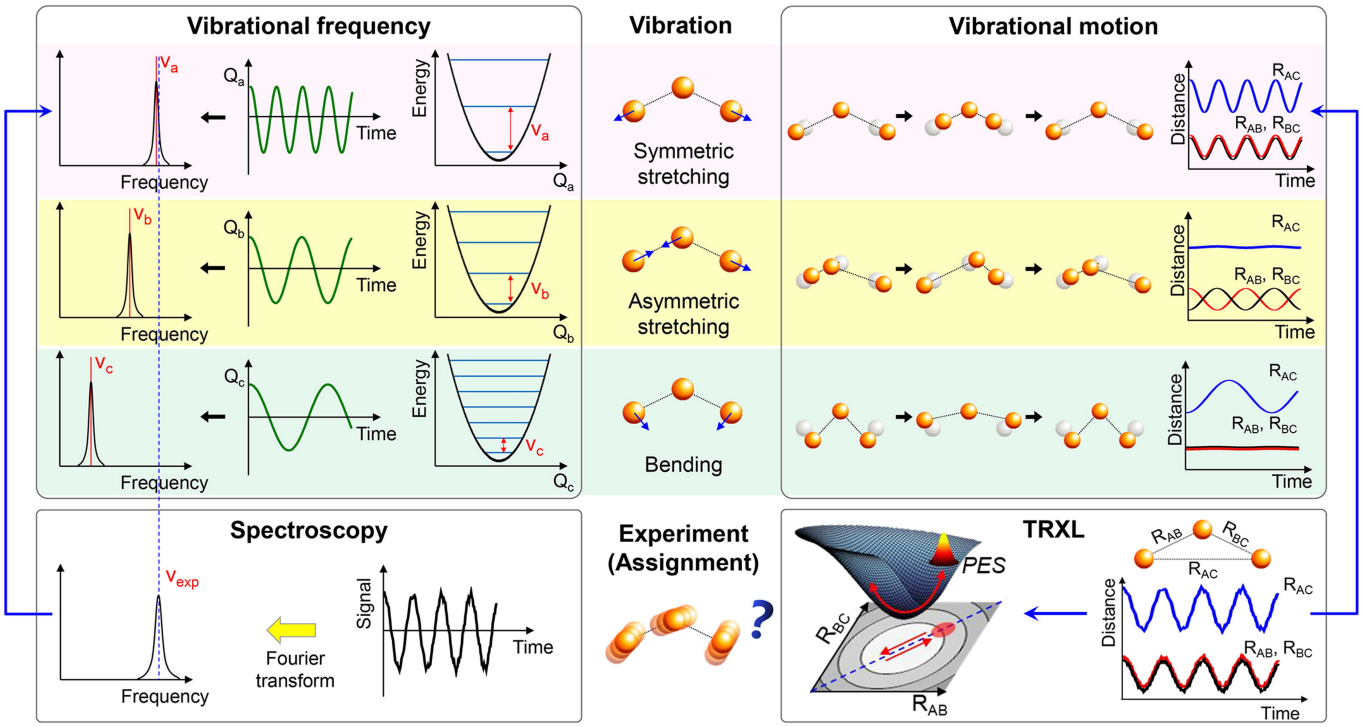
Extended Data Fig. 1 | Schematic of photoinduced bond formation in $[Au(CN)_2^-]_3$. Upon laser excitation (with energy represented by $h\nu$), wavepackets are created in both of the ground and excited states. The excited-state wavepacket in the T_1' state is prepared in the FC region after the ultrafast intersystem crossing from the initially excited singlet state (S_1) to a triplet excited state (T_1'). The excited-state wavepacket created in the FC region should move towards the equilibrium structure of T_1', which has two equivalent covalent Au–Au bonds between adjacent gold atoms (right inset, yellow spheres; blue and white spheres denote N and C atoms, respectively). The trajectory of the wavepacket from the FC region to the equilibrium structure of T_1' eventually determines the reaction trajectories of the ultrafast bond formation and hints towards its reaction mechanism. Three candidate reaction mechanisms of bond formation (paths 1, 2 and 3), described in the text, are represented by blue arrows on the nuclear coordinates of R_{AB} versus R_{BC}. In short, path 2 represents a concerted bond formation mechanism and path 1 and path 3 represent asynchronous bond formation mechanism. Path 1 and path 3 are distinct, depending on which bond is formed first between the A–B pair and the B–C pair. The initial motion of the excited-state wavepacket affects the initial motion of the ground-state wavepacket in the S_0 state, because impulsive Raman scattering generating the ground-state wavepacket can occur non-impulsively, owing to the finite pulse duration (~100 fs), as described in Supplementary Information. After the initial motions of the wavepackets in the ground and excited states, the wavepackets oscillate around their equilibrium structures.



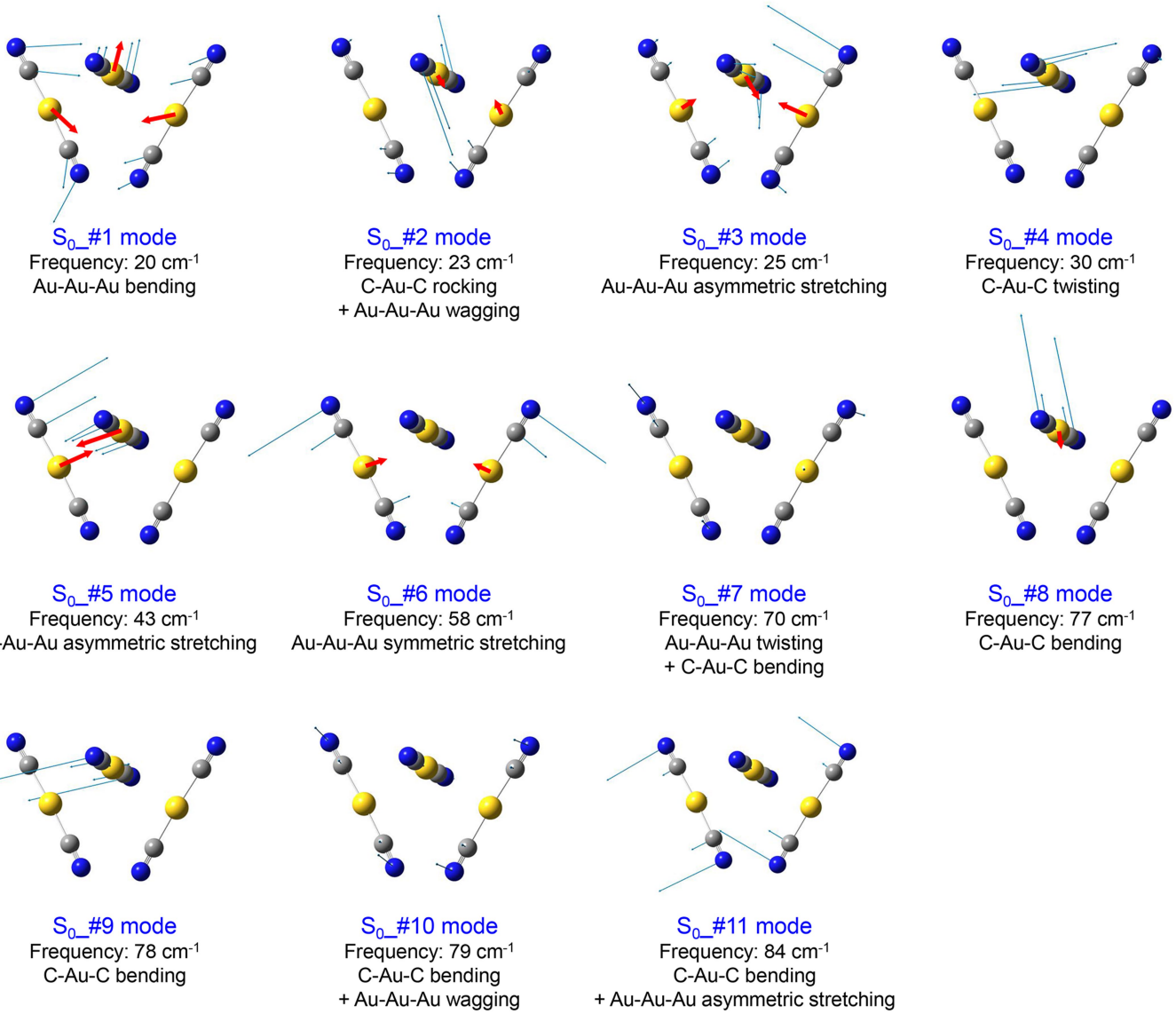
Extended Data Fig. 2 | TRXL data of $[Au(CN)_2^-]_3$ measured at PAL-XFEL and SACLA. a, **b**, Time-resolved difference scattering curves, $q\Delta S(q, t)$, of $[Au(CN)_2^-]_3$ measured at PAL-XFEL (**a**) and SACLA (**b**). **c**, The first and second RSVs (black and blue squares, respectively) obtained from the SVD analysis of $q\Delta S(q, t)$ measured at PAL-XFEL and their fits (red lines) using an exponential 1.1-ps time constant convoluted with an IRF with a FWHM of 170 fs. **d**, The first five RSVs resulting from the SVD analysis on the data measured at PAL-XFEL, multiplied by their corresponding singular values s_1, s_2, s_3, s_4 and s_5.



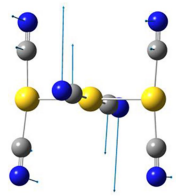
Extended Data Fig. 3 | Results of the structural analysis using residual difference scattering curves. a, Experimental residual difference scattering curves, $q\Delta S_{\mathrm{residual}}(q,t)$, measured from −1,040 fs to 2,235 fs. **b**–**e**, Theoretical fits of $q\Delta S_{\mathrm{residual}}(q,t)$ obtained from the structural analyses considering wavepacket motions in the S_0 state (**b**), the T_1' state (**c**), the T_1 state (**d**), or both the S_0 and T_1' states (**e**). Only the last analysis (using S_0 and T_1') gives a satisfactory fit quality.



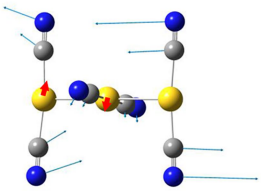
Extended Data Fig. 4 | Assignment of vibrational modes using vibrational frequencies and vibrational motions. Each vibrational normal mode has a specific structural motion with a characteristic frequency. For example, a simple nonlinear triatomic molecule has three vibrational modes named after specific structural motions: symmetric stretching, asymmetric stretching and bending. The characteristic frequency v_n of a vibrational mode vibrating along a normal coordinate Q_n corresponds to the energy gap between adjacent vibrational states of each mode, where $n = \{a, b, c\}$ for symmetric stretching, asymmetric stretching and bending, respectively. Vibrational frequencies are routinely measured by static or time-resolved spectroscopy that can probe vibrational transitions via infrared absorption or Raman scattering. Atomic motions themselves are not directly detected by spectroscopy, and thus the assignment of the observed frequencies to specific vibrational modes requires quantum chemical calculations that provide the connection between the vibrational frequencies and their corresponding atomic motions. By comparing the vibrational frequencies determined from experiment (v_{exp}) and quantum chemical calculation, the measured vibrational frequency can be assigned to a specific normal mode. Direct characterization of vibrational motions requires a tool with structural sensitivity, for example TRXL, as presented in this work. In a TRXL measurement, photoexcitation with a coherent optical laser pulse creates vibrational wavepackets of certain vibrational modes, and scattering of an X-ray pulse directly probes the resultant time-dependent structural changes that are characteristic of the activated vibrational modes—such as the temporal changes of the interatomic distances (R_{AB}, R_{BC} and R_{AC}) in $[Au(CN)_2^-]_3$. On the basis of direct information of both vibrational motions and vibrational frequencies obtained with TRXL, vibrational assignments can be made more accurately, and even the locations of vibrational wavepackets and the trajectories of their motions in multidimensional nuclear coordinates can be determined.



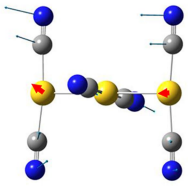
Extended Data Fig. 5 | Normal modes of the S_0 state. Normal modes of the S_0 state with frequencies in a range from 20 cm^{-1} to 170 cm^{-1}, obtained from DFT calculations. The frequency and atomic motions of each normal mode are shown. Displacement vectors of each normal mode are indicated by red arrows for the Au atoms (yellow) and blue arrows for the other atoms (C, grey spheres; N, blue spheres).



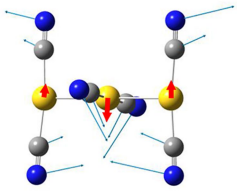
T_1_#1 mode
Frequency: 24 cm^{-1}
C-Au-C twisting



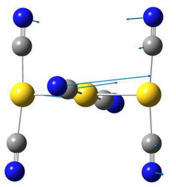
T_1_#2 mode
Frequency: 25 cm^{-1}
C-Au-C rocking
+ Au-Au-Au bending



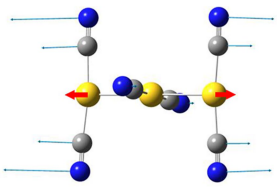
T_1_#3 mode
Frequency: 33 cm^{-1}
Au-Au-Au bending



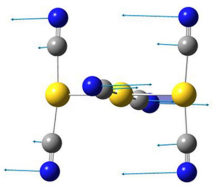
T_1_#4 mode
Frequency: 53 cm^{-1}
Au-Au-Au bending
+ C-Au-C rocking



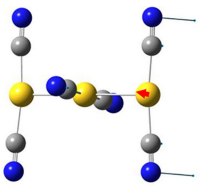
T_1_#5 mode
Frequency: 58 cm^{-1}
C-Au-C twisting



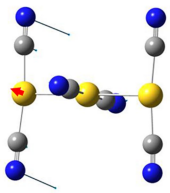
T_1_#6 mode
Frequency: 63 cm^{-1}
Au-Au-Au symmetric stretching



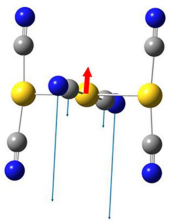
T_1_#7 mode
Frequency: 75 cm^{-1}
C-Au-C bending



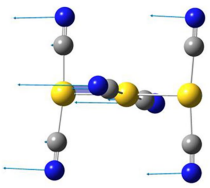
T_1_#8 mode
Frequency: 79 cm^{-1}
C-Au-C bending



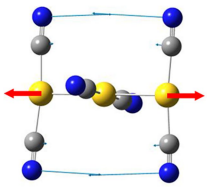
T_1_#9 mode
Frequency: 81 cm^{-1}
C-Au-C bending



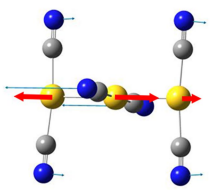
T_1_#10 mode
Frequency: 86 cm^{-1}
C-Au-C bending



T_1_#11 mode
Frequency: 87 cm^{-1}
C-Au-C bending

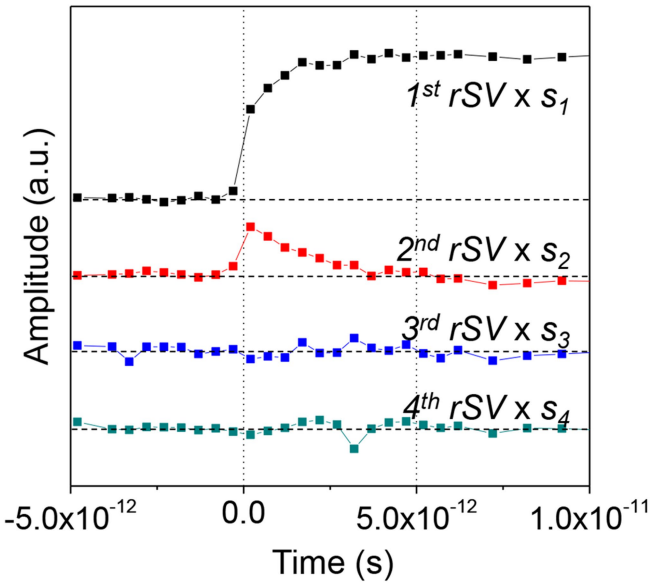


T_1_#12 mode
Frequency: 92 cm^{-1}
Au-Au-Au symmetric stretching

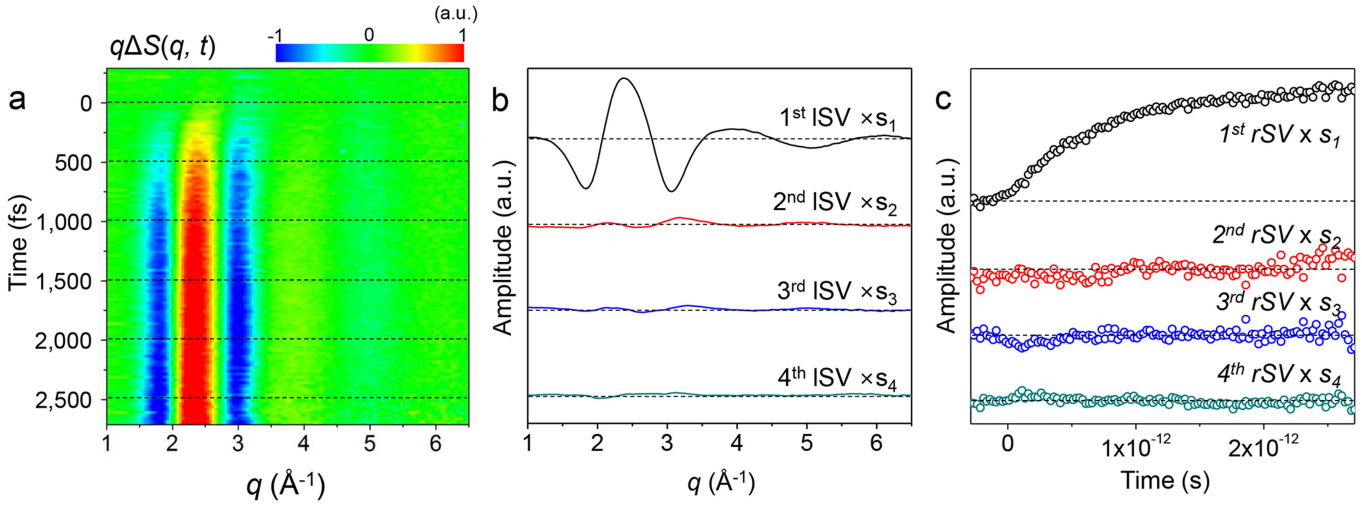


T_1_#13 mode
Frequency: 117 cm^{-1}
Au-Au-Au asymmetric stretching

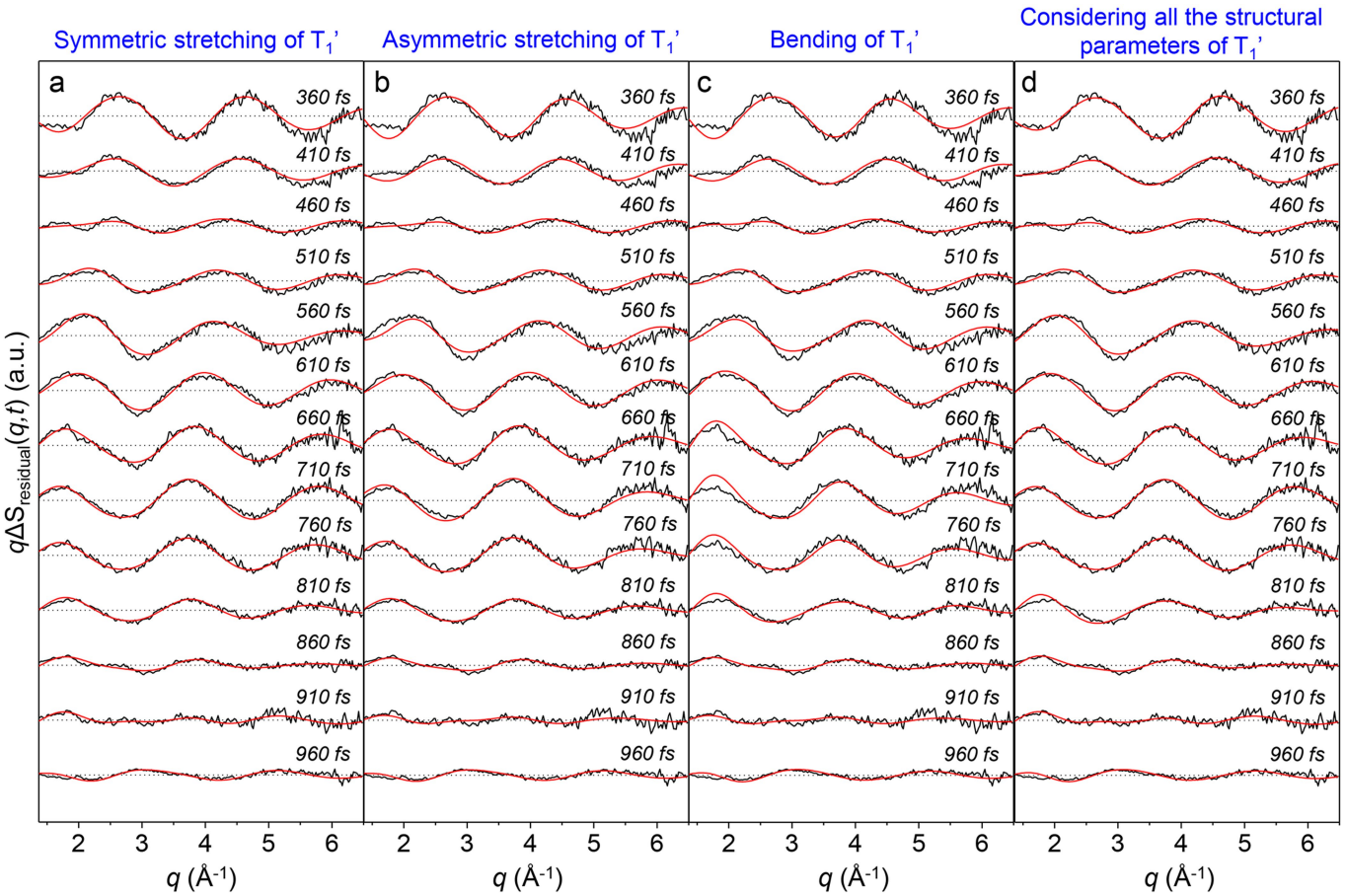
Extended Data Fig. 6 | Normal modes of the T_1 state. Normal modes of the T_1 state, as in Extended Data Fig. 5.



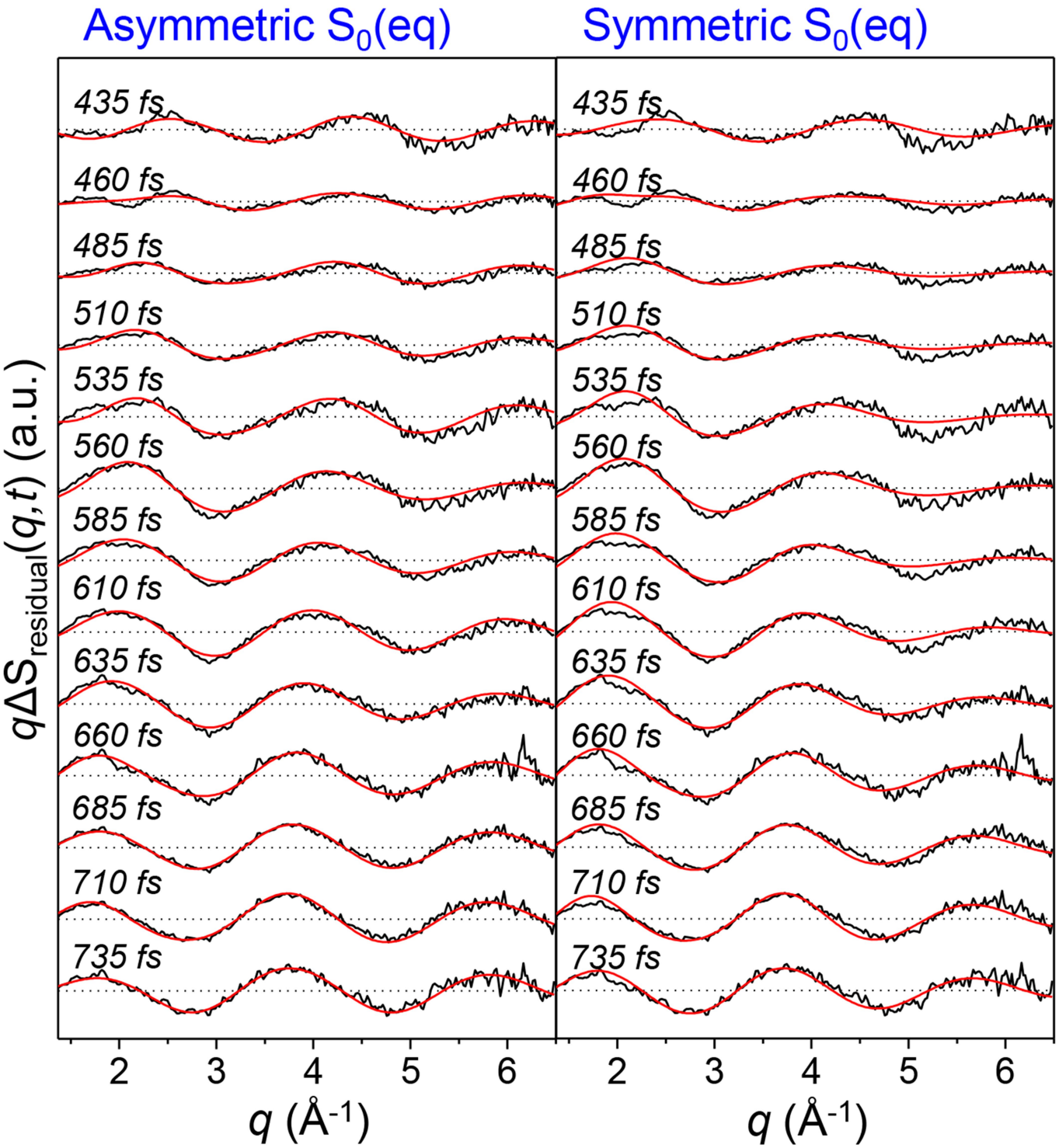
Extended Data Fig. 7 | SVD analysis on the TRXL data of $[Au(CN)_2^-]_3$ measured in the previous TRXL study. See ref. [9]. Shown are the first four RSVs multiplied by their corresponding singular values.



Extended Data Fig. 8 | Solvent heating contribution to the TRXL signal. **a**, Experimental difference scattering curves, $q\Delta S(q)$, of $FeCl_3$ solution measured at time delays from −740 fs to 2,260 fs. **b**, The first four LSVs multiplied by their corresponding singular values. **c**, The first four RSVs multiplied by their corresponding singular values. Only the first LSV and RSV contain meaningful signals, and so a single difference scattering curve (that is, the first LSV) accounts for the contribution to solvent heating on the scattering data measured with a water solvent. In the structural analysis, the first LSV was used as a scattering intensity change upon increase in temperature of the water solvent, $\Delta S_{heat}(q)$.



Extended Data Fig. 9 | Structural analysis of the three types of vibrational motions of the T_1 state. Experimental residual difference scattering curves (black lines) at several time delays after 360 fs and their theoretical fits (red lines), obtained from the structural analysis. **a**–**c**, For the structural analysis, vibrations of T_1' at later times (>360 fs) were classified into three types of vibrational motions (symmetric stretching, asymmetric stretching and bending). We performed the structural analysis for each of the three cases considering symmetric stretching motions (**a**), asymmetric stretching motions (**b**) and bending motions (**c**) of T_1'. **d**, For comparison, all the three structural parameters of T_1' were used for the structural analysis.



Extended Data Fig. 10 | Structural analysis using asymmetric and symmetric structures of S_0. Experimental residual difference scattering curves (black lines) at selected time delays and their theoretical fits (red lines) obtained from the structural refinements considering the asymmetric bent S_0 structure (left) or symmetric bent S_0 structure (right). The asymmetric case gives superior fit qualities compared to the symmetric case, indicating that the equilibrium structure of S_0 is asymmetric and bent.

Article

Variable water input controls evolution of the Lesser Antilles volcanic arc

https://doi.org/10.1038/s41586-020-2407-5

Received: 14 August 2019

Accepted: 26 March 2020

Published online: 24 June 2020

Check for updates

George F. Cooper[1,6]✉, Colin G. Macpherson[2], Jon D. Blundy[1], Benjamin Maunder[3], Robert W. Allen[3], Saskia Goes[3], Jenny S. Collier[3], Lidong Bie[5], Nicholas Harmon[4], Stephen P. Hicks[3], Alexander A. Iveson[2], Julie Prytulak[2], Andreas Rietbrock[5], Catherine A. Rychert[4], Jon P. Davidson[2] & the VoiLA team*

Oceanic lithosphere carries volatiles, notably water, into the mantle through subduction at convergent plate boundaries. This subducted water exercises control on the production of magma, earthquakes, formation of continental crust and mineral resources. Identifying different potential fluid sources (sediments, crust and mantle lithosphere) and tracing fluids from their release to the surface has proved challenging[1]. Atlantic subduction zones are a valuable endmember when studying this deep water cycle because hydration in Atlantic lithosphere, produced by slow spreading, is expected to be highly non-uniform[2]. Here, as part of a multi-disciplinary project in the Lesser Antilles volcanic arc[3], we studied boron trace element and isotopic fingerprints of melt inclusions. These reveal that serpentine—that is, hydrated mantle rather than crust or sediments—is a dominant supplier of subducted water to the central arc. This serpentine is most likely to reside in a set of major fracture zones subducted beneath the central arc over approximately the past ten million years. The current dehydration of these fracture zones coincides with the current locations of the highest rates of earthquakes and prominent low shear velocities, whereas the preceding history of dehydration is consistent with the locations of higher volcanic productivity and thicker arc crust. These combined geochemical and geophysical data indicate that the structure and hydration of the subducted plate are directly connected to the evolution of the arc and its associated seismic and volcanic hazards.

The 750-km-long Lesser Antilles volcanic arc (LAA), located along the eastern margin of the Caribbean Plate, is the result of slow (1–2 cm per year) westward subduction of Atlantic and proto-Caribbean oceanic lithosphere (Fig. 1). Water hosted in hydrous phases within the subducting plate will be released as the slab sinks into the mantle and warms up. As the water migrates out of the slab, the stress on faults is reduced, causing earthquakes. At the same time, the addition of water to the overlying mantle wedge reduces the solidus temperature, which may enhance melting. LAA magma production rates lie at the lower end of the global range, probably owing to the low convergence rates, and are very unevenly distributed, being greatest in the centre of the arc (Dominica and Guadeloupe)[4]. The LAA also displays notable along-arc variations in geochemistry, volcanic activity, crustal structure and seismicity[5–8]. Subducting plate velocity and age are often held responsible for variations in convergent margin behaviour[9] but are unlikely to have first-order influence on lateral variations within the LAA as neither vary greatly along-strike. Instead, variations in LAA magmatism and seismicity have been proposed to reflect a combination of (1) a strong north to south increase in sediment input[10]; (2) subduction of bathymetric ridges below the central arc[11], which may enhance plate stress and coupling; and/or (3) subduction of strongly hydrated fracture zones[12] at several locations along arc (Fig. 1).

Current plate reconstructions[13] show the northern LAA to be underlain by ~90-Myr-old subducted lithosphere that formed at the equatorial Mid-Atlantic Ridge and includes the Marathon and Mercurius fracture zones (Fig. 1), whereas beneath the southern LAA, the subducted lithosphere is up to 120 Myr old and formed at the, now fully subducted, proto-Caribbean mid-ocean ridge. The seafloor spreading rates were slow in both cases. The boundary between the two seafloor-spreading domains is clearly visible in both bathymetric and gravity data, projecting from the Demerara Plateau towards the central islands before becoming obscured by the accretionary prism around Barbados (Fig. 1; Extended Data Fig. 1).

Hydration of lithosphere formed by intermediate or fast spreading occurs mainly in the mafic crust through faults that form as the plate bends into the trench. By contrast, slow spreading produces highly tectonized oceanic lithosphere with relatively thin mafic crust, pronounced faults and sections of upper-mantle material exposed at the seafloor[14]. The transform faults at slow-spreading ridges, which manifest as fracture zones in mature oceanic crust, are more seismically

[1]School of Earth Sciences, University of Bristol, Bristol, UK. [2]Department of Earth Sciences, Durham University, Durham, UK. [3]Department of Earth Science and Engineering, Imperial College London, London, UK. [4]University of Southampton, National Oceanography Centre, Southampton, UK. [5]Geophysical Institute (GPI), Karlsruhe Institute of Technology, Karlsruhe, Germany. [6]Present address: School of Earth and Ocean Sciences, Cardiff University, Cardiff, UK. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: CooperG3@cardiff.ac.uk

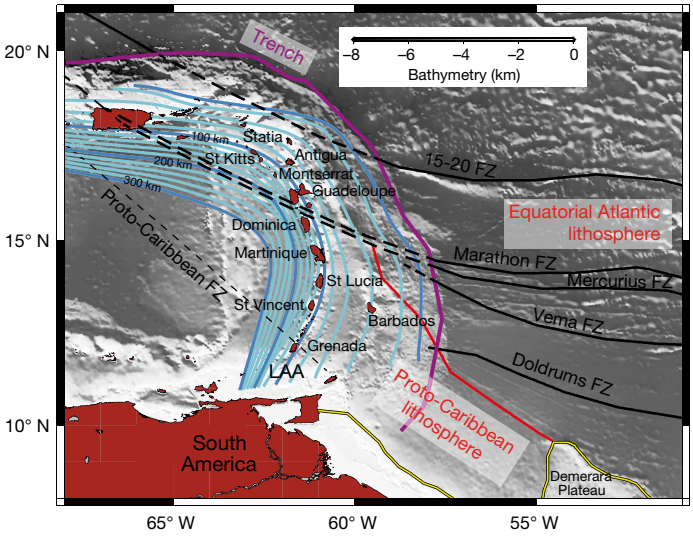


Fig. 1 | Bathymetric map of the study area, showing the islands of the LAA. Map shows locations of the trench (purple line), oceanic fracture zones (FZ; black lines, dashed where subducted), boundary between the proto-Caribbean and equatorial Atlantic seafloor (red line), and South American continent–ocean boundary (yellow line). Proto-Caribbean fracture zones have fully subducted; the likely location of a single one, required by basin geometry, is shown as a light dashed line. The bathymetric contrast between the northern and southern forearc is due to a strong difference in sediment thickness (from a few kilometres in the north to >15 km in the Barbados accretionary prism). Depth contours of the slab below the LAA are shown every 20 km (light blue lines) and every 100 km (dark blue lines). See Methods and Extended Data Figs. 1 and 2 for further details.

active and penetrate to greater depths than lithosphere formed by faster spreading[15]. These large-scale faults provide pathways for seawater and low/medium-temperature alteration including hydration of the mantle mineral olivine to serpentine[16]. Serpentine, in the form of antigorite, can hold up to 13 wt% structural water, at least double the water capacity of hydrated mafic crust. Thus, subduction of serpentinized mantle lithosphere has the potential to supply substantial volumes of fluid to magmatic arcs. To evaluate along-arc variations of slab-derived fluid sources (for example, sediment, oceanic crust, or serpentinized mantle lithosphere), we measured trace element concentrations and boron isotope ratios of melt inclusions in arc lavas along the entire LAA. To investigate how fluids influence arc magma genesis and evolution, we compare these geochemical proxies for slab-derived fluids with newly acquired geophysical data[3], and with the predicted positions of subducted fracture zones and the proto-Caribbean/Equatorial Atlantic plate boundary below the arc at different times.

In subduction zone magmas, boron and its isotopes trace contributions from fluids released by the subducting plate[17,18]. Boron is fluid-mobile, and a high ratio of boron to fluid-immobile elements, such as Ti, Nb or Zr, in arc magmas suggests that boron is principally supplied by subducting-plate fluids[19]. Serpentine-derived boron is enriched in ^{11}B compared with ^{10}B, producing distinctively elevated $\delta^{11}B$ values of +7‰ to +20‰ (ref. [17]) ($\delta^{11}B = (((^{11}B/^{10}B)_{sample}/(^{11}B/^{10}B)_{standard} - 1))$. As a result, arc magmas produced through mantle melting induced by serpentine-derived fluids have significantly higher $\delta^{11}B$ values (up to +18‰)[20] than mantle from mid-ocean-ridge basalt sources (−7.1 ± 0.9‰)[21]. Fluids derived from subducted sediments also have a different distinct chemical signature[22]. Sediments in ocean drill cores east of the LAA contain terrigenous turbidites, pelagic clays and ashy siliceous clays[23]. Although these sediments are enriched in boron (50–160 ppm B), they have significantly lower $\delta^{11}B$ values (approximately −15‰ to +5‰)[21] than serpentine-derived fluids at sub-arc depths[24].

Using secondary-ion mass spectrometry (SIMS), we measured 198 glassy, clinopyroxene-hosted melt inclusions for volatiles (H_2O, CO_2) and trace elements, of which 92 were further analysed for boron isotopic composition. The analysed melt inclusions are from fresh volcanic deposits assumed to be much less than 1 Myr old, and range from low-MgO, high-alumina basalt (MgO = 1.8–3.5 wt%, Al_2O_3 = 15.3–19.1 wt%) to rhyolite (≤78 wt% SiO_2; Fig. 2). All of these compositions have undergone some level of magmatic differentiation in the shallow crust, so none can be considered primary; however, the boron isotopic signature is largely determined by the source rather than subsequent differentiation processes[25,26]. We supplemented our data set with all

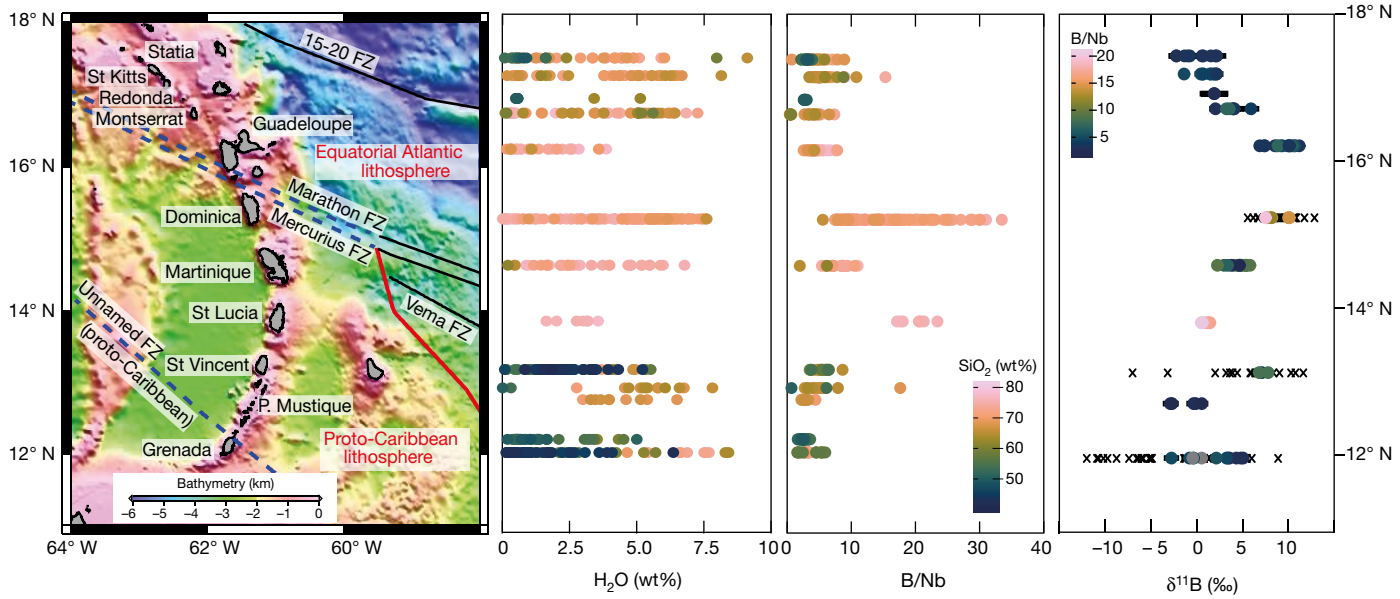


Fig. 2 | Bathymetric map of the LAA compared with water, B/Nb ratios and $\delta^{11}B$ of melt inclusions in lavas. H_2O (this study and compiled published values) and B/Nb symbols are coloured by the SiO_2 wt% of melt inclusions, as an indicator of magmatic differentiation. $\delta^{11}B$ symbols are coloured by B/Nb as an indicator of fluid addition. Previously published boron isotope ratios from melt inclusions[33–35] are shown as crosses. Error bars on $\delta^{11}B$ values represent propagated 1σ uncertainties and are typically less than ±1‰.

previously published LAA melt inclusion analyses (n > 1,000) available from the GEOROC database.

LAA melt inclusions are characterized by dissolved water contents of up to 9.1 wt% H_2O, with a large range for individual islands (Fig. 2). However, water contents of melt inclusions are affected by differentiation processes during crustal storage and thus are a poor proxy for primary magmatic water contents. Water content will increase in a melt undergoing undersaturated crystallization, remain constant under water-saturated conditions and be lost from melt during late-stage degassing. Further modification of water in melt inclusions can occur because of post-entrapment crystallization and/or diffusive water loss. Ratios of fluid-mobile to fluid-immobile trace elements, such as B/Nb (Fig. 2), are more reliable indicators of the contribution of fluids, as both elements behave similarly during melting and magmatic differentiation. Our data shows high ratios of B/Nb in the central arc, which most probably reflect a particularly fluid-rich and B-rich magmatic source.

The new $\delta^{11}B$ values for LAA melt inclusions vary from −2.8‰ to +11.2‰ (Fig. 2), which spans much of the global arc range (−9‰ to +16‰)[17]. Melt inclusions with the highest $\delta^{11}B$ values are from the central arc (islands of Guadeloupe and Dominica; Fig. 2). Variation of $\delta^{11}B$ within each volcanic centre is unlikely to be due to crustal differentiation because there are no systematic trends in $\delta^{11}B$ with indicators of differentiation (for example, SiO_2 and Rb/Sr, Extended Data Fig. 3). This is consistent with prior findings that fractional crystallization has negligible effect on melt $\delta^{11}B$ values[25,26]. Crustal assimilation during open-system differentiation may also modify $\delta^{11}B$ and B/Nb, but inputs from this source probably have a similar isotopic and geochemical composition to AOC and sediment[22]. Assimilation of LAA crust would lower melt $\delta^{11}B$ values during differentiation, a trend that is not observed in our data (Extended Data Fig. 3). Although there is a range of melt-inclusion $\delta^{11}B$ values within each single volcanic centre (for example, 3.5‰ in Martinique), there are clear $\delta^{11}B$ differences between neighbouring volcanic centres with similar major element chemistry. Therefore, we interpret the distinct $\delta^{11}B$ values in evolved melt inclusions at each island as a reflection of differences between the mantle source regions of each island, such that boron isotopes provide a robust tracer for the fluid source[18].

We interpret the $\delta^{11}B$ differences between islands and the systematic $\delta^{11}B$ change along the arc to result from variable involvement of fluids from two distinct sources: (1) altered oceanic crust (AOC) and sediment; and (2) serpentine dehydration (Fig. 3). In the central portion of the arc, melt inclusions from Guadeloupe and Dominica have $\delta^{11}B$ values significantly greater than +5‰. Of the available sources, only fluid with >60% contribution from serpentine dehydration has the capacity to generate this isotopic signature (Fig. 3). The lower $\delta^{11}B$ values found in the north and south of the arc can be attributed primarily to fluid released by dehydration of AOC and sediment (Fig. 3). However, there is no simple relationship between $\delta^{11}B$ and indicators of varying volume of fluid addition (for example, B/Be and Nb/B; Extended Data Fig. 3; Fig. 3). In contrast to Guadeloupe and Dominica, St Lucia melt inclusions from this study have a high net fluid contribution based on the Nb/B values, but we estimate that <30% of this originates from serpentine. Therefore, the total volume of fluid is decoupled from the proportion of different sources from which each fluid is derived. In the north and south of the arc, with the exception of St Vincent, the proportion of fluid derived from serpentine is lower than in the central arc. Based on boron isotopes, it is not possible to distinguish whether the serpentinite fluids are derived from the slab or from recycled forearc material[20,27]. However, a peak in seismicity occurs in the central arc at the depths where models predict dehydration of peridotite in the slab (120–160 km)[9,28]. In conjunction with the abundance of serpentinized peridotite expected in lithosphere formed by slow spreading[14,29], this provides an argument for slab-hosted serpentine being the main deliverer of fluid to the LAA mantle wedge.

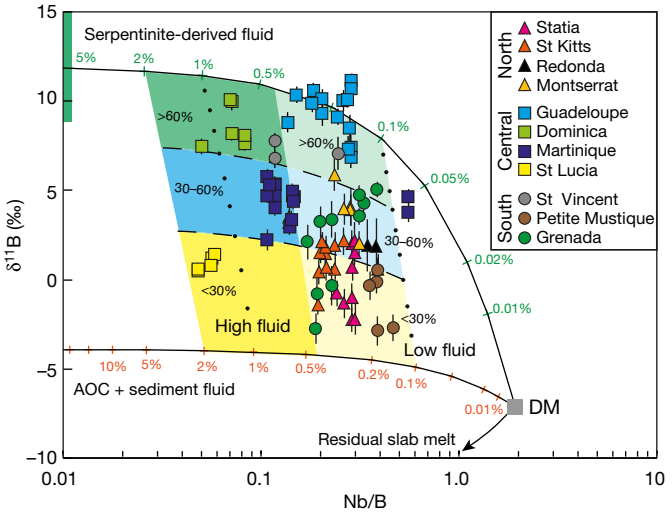


Fig. 3 | Melt inclusion Nb/B versus $\delta^{11}B$ for LAA magmas from this study. Mixing model (black lines) shows contamination of depleted mantle (DM, grey square) by fluid derived from serpentinite and from altered oceanic crust (AOC) plus sediment-derived fluids at 120 km depth. Green bar represents global serpentinite range. Red and green numbers represent the percentage by mass of fluid from the two sources added to the mantle. Inputs for the model are detailed in the Methods. Dotted lines indicate composite fluids formed by mixing between (0.1% and 1% mass) fluids from the two discrete sources. Shading indicates >60% (green), 30–60% (blue) and <30% (yellow) contribution from subducted serpentinite. Darker and lighter shaded areas represent domains referred to in text as 'high' and 'low' fluid contributions, respectively. Only samples measured in this study are plotted (n = 92). Error bars on $\delta^{11}B$ values represent propagated 1σ uncertainties and are smaller than symbol size where absent. All 1σ uncertainties are typically less than ±1‰.

We compared our geochemical results to a range of independent observations that may be expressions of fluid release (Fig. 4). As these observations sample different parts of the subduction system in space and time, we modelled expected excess hydration (that is, fluid derived from fractures zones) to the arc over the past 25 Myr (Fig. 4b), assuming that the known fracture zones and plate boundary between the proto-Caribbean and Atlantic bring extra water in the form of serpentine (see Methods).

If higher recent fluid fluxes below the arc were to cause an increase in magmas production, then we might expect to see boron isotope ratios (Fig. 4a) and/or intra-slab seismicity rates[30] correlate with volcanic production rates[4] (Fig. 4e, f). Slab seismicity is often attributed to dehydration embrittlement[31], and the depth to which seismicity extends[30] is consistent with the extent of the serpentinite stability field predicted for the convergence rates and ages of LAA subduction. Our data show a peak in boron isotopes, intra-slab seismicity rates and volcanic production rates around Dominica, and this is where our forward models (Fig. 4b) predict a peak in dehydration from 0 to 2 million years ago (Ma) due to the subduction of the Marathon and Mercurius fracture zones. Therefore, our data indicate that enhanced fluid fluxing of the mantle wedge is associated with higher magma production in the LAA. However, it is not possible to quantify how much of the excess fluid release contributes to enhanced flux melting versus enhanced decompressional melting.

High ratios of small to large earthquakes (high b-values) on the plate interface and forearc[12] (Fig. 4c), as well as low shear-wave velocities (4.3 ± 0.05 km s^{-1}) at 50 km depth (Fig. 4d, derived from Rayleigh waves recorded during the VoiLA seismic experiment[3]; see Methods) could reflect excess dehydration at shallower depths. High b-values are commonly attributed to seismogenic failure at lower stresses due to higher pore fluid pressures, while shear velocity anomalies of around 9% could

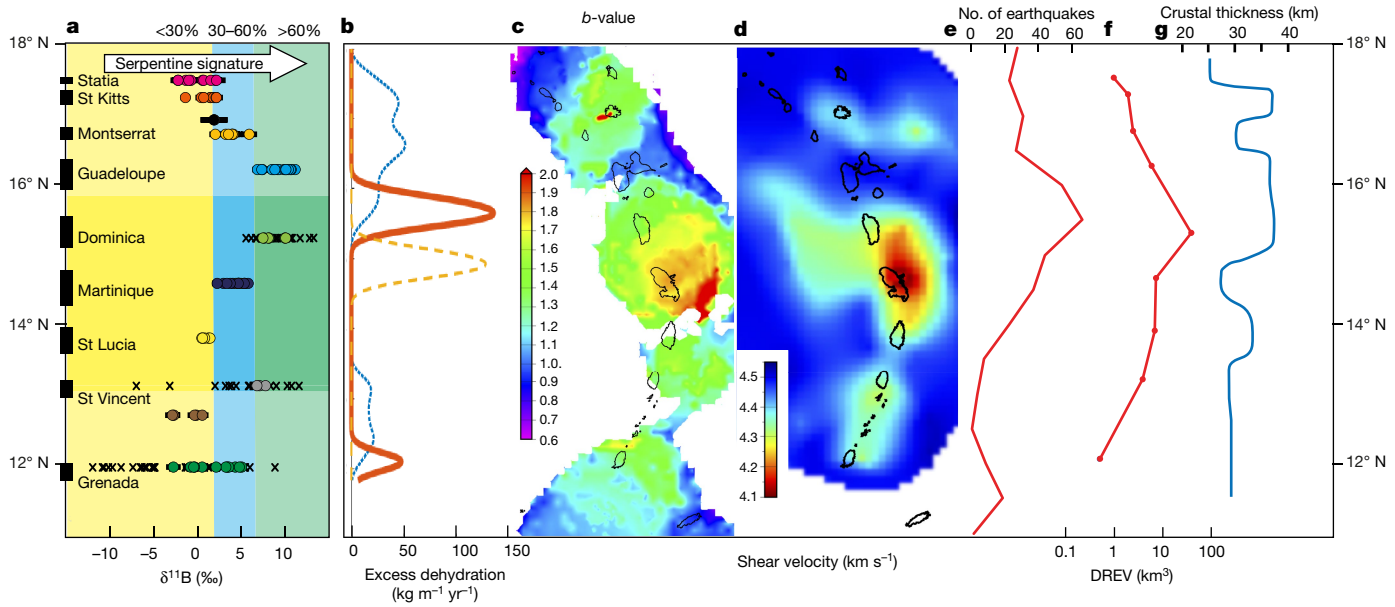


Fig. 4 | Summary of along-arc geochemical and geophysical data. a, Boron isotope ratios of melt inclusions with latitude in the LAA (data symbols coloured as in Fig. 3; previously published data[33–35] shown by crosses). Light and dark coloured shaded areas correspond to those in Fig. 3. **b**, Modelled sub-arc excess (that is, fracture-zone-associated) dehydration averaged over the past 2 Myr (solid red line for fluids released below the arc, dashed yellow line below the forearc) and 25 Myr (dotted blue line, below the arc) based on plate reconstruction and slab geometry; see Methods. The excess dehydration represents the rate of water released from the slab above uniform background level. **c**, *b*-value distribution (relative frequency of small versus large events below the forearc)[12]. **d**, Shear-wave velocity from teleseismic Rayleigh waves at 50 km depth, with main anomalies below the forearc. **e**, Local seismicity in the subducting plate[30]. **f**, Volcanic production rates over the past 100 kyr as dense-rock-equivalent volumes (DREV)[3]. Red lines in **e** and **f** correspond to excess dehydration over the past 2 Myr (red line in **b**). **g**, Crustal thickness below the arc from receiver functions[7] (blue line). The modelled trends compare well with the main anomalies in data sensitive to recent fluid release below the forearc (**c**, **d**), below the arc (**e**, **f**) and over the past 25 Myr (**g**).

correspond to about 1.1 vol% of fluids and associated melts[32]. Shear velocities and *b*-values are characterized by a prominent maximum and minimum, respectively, in the region around Martinique, that is, displaced southward from the peak in boron isotopes. Owing to the obliquity of the fracture zones to the trench, excess forearc dehydration (derived from shallower slab depths) is expected to occur further to the south than dehydration below the arc, coincident with the *b*-value and shear velocity peaks (Fig. 4b).

Finally, there are systematic variations in crustal thickness along the arc[7], with thicknesses of around 35 km north of Martinique and around 30 km in the south. These reflect a long-term integrated variation in magma productivity. When we consider the excess dehydration over the age of the present arc (around 25 Myr), the position of Marathon–Mercurius fracture zone subduction has shifted from the north near St Kitts to Dominica today, so a larger crustal thickness would be expected along the whole northern arc, as observed. Again, however, we cannot constrain the relative role of decompression melting in this magma production.

None of the other Atlantic fracture zones has contributed to dehydration below the arc. The Fifteen-Twenty (15-20) fracture zone has not subducted deep enough (but higher *b*-values and lower shear-wave velocities in the forearc near Antigua in Fig. 4 could, given the spatial resolution of these measurements, indicate shallow fluid release from it). Other Atlantic fracture zones have yet to reach the trench. It is likely that there were fracture zones in the proto-Caribbean oceanic lithosphere, but their location is uncertain. We included in our model a single, large-offset fracture zone at the location required to fit the basin geometry between the Bahamas Bank and Demerara Rise (Fig. 1; see Methods). This yields a small peak in excess dehydration in the southernmost arc. Thus, within the uncertainties, proto-Caribbean fracture zones could explain the increases in δ^{11}B and *b*-values and the decrease in shear velocities around St Vincent and Grenada.

Given the geological complexity of subduction systems, our new geochemical and geophysical expressions of fluids along the LAA show remarkable coherence with the predicted history of fluid release from fracture zones in the subducting plate at different locations in the system and over different temporal windows. Furthermore, the high boron contents and elevated δ^{11}B signature of melt inclusions in magmas from the central segment of the arc are unambiguous indicators of dehydration of subducted serpentine, which is expected to be one of the main minerals formed in fracture zone hydration. Therefore, our observations provide strong evidence that a heterogeneous distribution of serpentine in subducting mantle lithosphere exerts a primary control on along-arc variations in mantle wedge hydration and seismicity and may also influence the crustal structure and magmatic productivity of volcanic arcs.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2407-5.

1. Hacker, B. R. H_2O subduction beyond arcs. *Geochem. Geophys. Geosystems* **9**, Q03001 (2008).
2. Grevemeyer, I., Ranero, C. R. & Ivandic, M. Structure of oceanic crust and serpentinization at subduction trenches. *Geosphere* **14**, 395–418 (2018).
3. Goes, S. et al. Project VoiLA: Volatile Recycling in the Lesser Antilles. *Eos* **100**, https://doi.org/10.1029/2019EO117309 (2019).
4. Wadge, G. Comparison of volcanic production rates and subduction rates in the Lesser Antilles and Central America. *Geology* **12**, 555–558 (1984).
5. Boynton, C. H., Westbrook, G. K., Bott, M. H. P. & Long, R. E. A seismic refraction investigation of crustal structure beneath the Lesser Antilles island arc. *Geophys. J. R. Astron. Soc.* **58**, 371–393 (1979).
6. Macdonald, R., Hawkesworth, C. J. & Heath, E. The Lesser Antilles volcanic chain: a study in arc magmatism. *Earth Sci. Rev.* **49**, 1–76 (2000).

7. Melekhova, E. et al. Lateral variation in crustal structure along the Lesser Antilles arc from petrology of crustal xenoliths and seismic receiver functions. *Earth Planet. Sci. Lett.* **516**, 12–24 (2019).
8. Hayes, G. P., McNamara, D. E., Seidman, L. & Roger, J. Quantifying potential earthquake and tsunami hazard in the Lesser Antilles subduction zone of the Caribbean region. *Geophys. J. Int.* **196**, 510–521 (2014).
9. van Keken, P. E., Hacker, B. R., Syracuse, E. M. & Abers, G. A. Subduction factory: 4. Depth-dependent flux of H_2O from subducting slabs worldwide. *J. Geophys. Res. Solid Earth* **116**, B01401 (2011).
10. Carpentier, M., Chauvel, C. & Mattielli, N. Pb–Nd isotopic constraints on sedimentary input into the Lesser Antilles arc system. *Earth Planet. Sci. Lett.* **272**, 199–211 (2008).
11. Bouysse, P. & Westercamp, D. Subduction of Atlantic aseismic ridges and Late Cenozoic evolution of the Lesser Antilles island arc. *Tectonophysics* **175**, 349–380 (1990).
12. Schlaphorst, D. et al. Water, oceanic fracture zones and the lubrication of subducting plate boundaries—insights from seismicity. *Geophys. J. Int.* **204**, 1405–1420 (2016).
13. Müller, R. D. et al. A global plate model including lithospheric deformation along major rifts and orogens since the Triassic. *Tectonics* **38**, 1884–1907 (2019).
14. Escartín, J. et al. Central role of detachment faults in accretion of slow-spreading oceanic lithosphere. *Nature* **455**, 790–794 (2008).
15. Manea, V. C., Leeman, W. P., Gerya, T., Manea, M. & Zhu, G. Subduction of fracture zones controls mantle melting and geochemical signature above slabs. *Nat. Commun.* **5**, 5095 (2014).
16. Bach, W. & Früh-Green, G. L. Alteration of the oceanic lithosphere and implications for seafloor processes. *Elements* **6**, 173–178 (2010).
17. De Hoog, J. C. M. & Savov, I. P. in *Boron Isotopes: The Fifth Element* (eds Marschall, H. & Foster, G.) 217–247 (Springer, 2018).
18. Leeman, W. P., Tonarini, S. & Turner, S. Boron isotope variations in Tonga–Kermadec–New Zealand arc lavas: implications for the origin of subduction components and mantle influences. *Geochem. Geophys. Geosyst.* **18**, 1126–1162 (2017).
19. Leeman, W. P. in *Subduction: Top to Bottom* (eds Bebout, G. E. et al.) 269–276 (AGU, 1996).
20. Tonarini, S., Leeman, W. P. & Leat, P. T. Subduction erosion of forearc mantle wedge implicated in the genesis of the South Sandwich Island (SSI) arc: evidence from boron isotope systematics. *Earth Planet. Sci. Lett.* **301**, 275–284 (2011).
21. Marschall, H. R. in *Boron Isotopes: The Fifth Element* (eds Marschall, H. & Foster, G.) 189–215 (Springer, 2018).
22. Bezard, R. et al. Assimilation of sediments embedded in the oceanic arc crust: myth or reality? *Earth Planet. Sci. Lett.* **395**, 51–60 (2014).
23. Plank, T. in *Treatise on Geochemistry* 2nd edn (eds Holland, H. D. & Turekian, K. K.) 607–629 (Elsevier, 2014).
24. Benton, L. D., Ryan, J. G. & Tera, F. Boron isotope systematics of slab fluids as inferred from a serpentine seamount, Mariana forearc. *Earth Planet. Sci. Lett.* **187**, 273–282 (2001).
25. Kaliwoda, M. et al. Boron and boron isotope systematics in the peralkaline Ilímaussaq intrusion (South Greenland) and its granitic country rocks: a record of magmatic and hydrothermal processes. *Lithos* **125**, 51–64 (2011).
26. Jones, R. E. et al. Temporal variations in the influence of the subducting slab on Central Andean arc magmas: evidence from boron isotope systematics. *Earth Planet. Sci. Lett.* **408**, 390–401 (2014).
27. McCaig, A. M. et al. No significant boron in the hydrated mantle of most subducting slabs. *Nat. Commun.* **9**, 4602–10 (2018).
28. Paulatto, M. et al. Dehydration of subducting slow-spread oceanic lithosphere in the Lesser Antilles. *Nat. Commun.* **8**, 15980 (2017).
29. Vils, F., Tonarini, S., Kalt, A. & Seitz, H.-M. Boron, lithium and strontium isotopes as tracers of seawater–serpentinite interaction at Mid-Atlantic Ridge, ODP Leg 209. *Earth Planet. Sci. Lett.* **286**, 414–425 (2009).
30. Bie, L. et al. Along-arc heterogeneity in local seismicity across the Lesser Antilles subduction zone from a dense ocean-bottom seismometer network. *Seismol. Res. Lett.* **91**, 237–247 (2020).
31. Kirby, S., Engdahl, R. E. & Denlinger, R. in *Subduction: Top to Bottom* (eds Bebout, G. E. et al.) 195–214 (AGU, 1996).
32. Hammond, W. C. & Humphreys, E. D. Upper mantle seismic wave velocity: effects of realistic partial melt geometries. *J. Geophys. Res. Solid Earth* **105**, 10975–10986 (2000).
33. Gurenko, A. A., Trumbull, R. B., Thomas, R. & Lindsay, J. M. A melt inclusion record of volatiles, trace elements and Li–B isotope variations in a single magma system from the Plat Pays volcanic complex, Dominica, Lesser Antilles. *J. Petrol.* **46**, 2495–2526 (2005).
34. Bouvier, A.-S., Métrich, N. & Deloule, E. Light elements, volatiles, and stable isotopes in basaltic melt inclusions from Grenada, Lesser Antilles: inferences for magma genesis. *Geochem. Geophys. Geosystems* **11**, Q09004 (2010).
35. Bouvier, A.-S., Manzini, M., Rose-Koga, E. F., Nichols, A. R. L. & Baumgartner, L. P. Tracing of Cl input into the sub-arc mantle through the combined analysis of B, O and Cl isotopes in melt inclusions. *Earth Planet. Sci. Lett.* **507**, 30–39 (2019).

The VoiLA team

George F. Cooper[1,7], Colin G. Macpherson[2], Jon D. Blundy[1], Benjamin Maunder[3], Robert W. Allen[3], Saskia Goes[3], Jenny S. Collier[3], Lidong Bie[5], Nicholas Harmon[4], Stephen P. Hicks[3], Andreas Rietbrock[5], Catherine A. Rychert[4], Jon P. Davidson[2], Richard G. Davy[3], Tim J. Henstock[4], Michael J. Kendall[2], David Schlaphorst[2], Jeroen van Hunen[2], Jamie J. Wilkinson[3,7] & Marjorie Wilson[8]

[7]Department of Earth Sciences, Natural History Museum, London, UK. [8]School of Earth and Environment, University of Leeds, Leeds, UK.

Methods

Geochemistry

Sample preparation. Crystals were separated from crushed and sieved scoria, pumice or lava. Picked crystals from the size fractions 0.5–1 mm and 1–2 mm were mounted on glass slides within 2.5-cm-diameter aluminium rings, back-filled with epoxy resin and polished to expose the centre of the crystals. Crystals were imaged under transmitted light to locate the most suitable glassy inclusions before further polishing to expose the maximum number of melt inclusions. All epoxy mounts were gold-coated before SIMS analysis.

Trace elements by SIMS. We measured concentrations of H_2O, CO_2 and trace elements in 198 melt inclusions using the Cameca IMS-4f at the NERC Edinburgh Ion Micro-Probe Facility (EIMF), over two sessions (October 2017 and January 2018). The IMS-4f instrument was run with a 15-kV (nominal) primary beam of O^- ions with a beam current of ~5 nA, resulting in a spot size at the sample surface of ~15 µm diameter. Positive secondary ions were extracted at 4.5 kV, using energy filtering with an energy window of 50 ± 25 eV (for CO_2 analysis) or 75 ± 25 eV (for all other elements). CO_2 measurements were performed first. Before each analysis, the sample was pre-sputtered using a primary beam raster of 20 µm for 4 min to reduce C backgrounds resulting from surface contamination. The isotopes $^{12}Mg^{2+}$, ^{12}C, ^{26}Mg and ^{30}Si were measured. Peak positions were verified at the start of each analysis. The background C signal was determined through analysis of the nominally C-free KL2-G glass standard. Following CO_2 analysis, H_2O and trace element concentrations were measured on the same analytical spot as the CO_2 analyses, using a secondary accelerating voltage of 4,500 V with 75-V offset and a 25-µm image field. The isotopes ^1H, ^7Li, ^{11}B, ^{19}F, ^{26}Mg, ^{35}Cl, ^{30}Si, ^{42}Ca, ^{44}Ca, ^{45}Sc, ^{47}Ti, ^{84}Sr, ^{85}Rb, ^{88}Sr, ^{89}Y, ^{90}Zr, ^{93}Nb, ^{133}Cs, ^{138}Ba, ^{139}La, ^{140}Ce and ^{149}Sm were measured. Calibration was carried out on a range of basaltic glass standards with 0–4 wt% H_2O, repeated throughout the day. Absolute element concentrations were calculated using the in-house JCION5 software and by normalizing the intensities to Si (as measured using ^{30}Si) which was determined by subsequent electron microprobe analysis. A summary of repeat analyses of GSD-1G and T1-G is presented in the Supplementary Data.

Electron microprobe. Following volatile and trace element analysis, we measured major elements using a Cameca SX100 electron microprobe (EPMA) at the University of Bristol, UK. The gold coat was removed, and samples were carbon-coated. Concentrations of SiO_2, TiO_2, Al_2O_3, Fe_2O_3, MnO, MgO, CaO, Na_2O, K_2O, P_2O_5, Cr_2O_3, SO_2 and Cl in glass were made with a 20-kV accelerating voltage, a 4-nA beam current and a 5-µm or 10-µm defocused beam to minimize alkali loss[36]. Major elements were calibrated using a range of synthetic oxide, mineral and metal standards.

Boron isotopes by SIMS. Before boron isotope analysis, crystals hosting the measured melt inclusions were cut out of the epoxy mounts and pressed into indium within 24-mm-diameter aluminium holders. This step reduced the total number of sample mounts and, as indium outgasses less than epoxy, reduces the time required to reach a suitable vacuum for analysis.

We measured boron isotopes (^{11}B and ^{10}B) in 92 melt inclusions using the Cameca IMS-1270 at the NERC EIMF in December 2018. Before analysis, the samples were cleaned and a gold coat was applied. Positive secondary ions of $^{10}B^+$ and $^{11}B^+$ were produced by sputtering the sample with a 5-nA, $^{16}O^{2-}$ primary beam with a net impact energy of 22 keV, focused using Köhler illumination to a spot size of ~25 µm. Secondary ions were extracted at 10 kV and counted by a single electron multiplier detector. No energy filtering was applied. Analyses were performed with a mass resolution ($M/\Delta M$) of ~2,400. Single analyses consisted of 50 measurement cycles of ^{10}B and ^{11}B signals, using counting times of 2 s. Instrumental fractionation was determined by using the reference materials GSD1-G, B6, GOR132-G, StHs6/80-G and BCR2-G, measured at the beginning, during and end of the session (Supplementary Data).

Boron mixing model

Element contents for AOC and sediment and serpentinite-derived fluids are from ref. [20]. Isotope ratios used for serpentinite fluids lie within the range of Atlantic peridotites[29,37–39]. Depleted mantle boron concentrations and isotope ratios are from ref. [40]; Nb concentrations are from ref. [41]. Values are presented in Extended Data Table 1. Composite fluids are produced by mixing the two most important endmembers in the Lesser Antilles (AOC + sediment and serpentinite-derived fluid).

Shear velocity

The ocean-bottom seismic data analysed in this study were collected during two cruises aboard the RRS *James Cook*[42,43]. We used vertical seismograms to measure the amplitude and phase of ambient noise cross-correlation function and teleseismic Rayleigh waves. The onshore and offshore data were corrected for instrument response, detrended and means removed before processing. The teleseismic data were further processed as detailed in ref. [9]. Measurements of Rayleigh wave dispersion and estimates of the amplitude at selected period were made using frequency-time analysis[44,45]. We measured dispersion from 11 s to 18 s period. We used up to 2,486 dispersion measurements from 93 events from teleseismic Rayleigh waves in the tomography.

Shear velocity tomography was performed in two steps: first we inverted the amplitude and phase data for the phase velocity maps[46–48], and then at each location in the phase velocity maps we inverted the one-dimensional shear velocity structure to generate a three-dimensional volume[46]. For the shear velocity inversion, we included the effects of the water column and sediment using a priori information; our initial crustal thickness was based on Airy isostasy across the region. The tomographic inversion subsequently solved for the best fitting crustal thickness as well as shear velocity.

Plate reconstruction and hydration modelling

Mapping the tectonic features. Our modelling of the subducted features below the Lesser Antilles is based upon the global plate reconstruction of ref. [13]. as implemented within the software G-Plates 2.1 (https://doi.org/10.1029/2018GC007584). In this reconstruction, the opening of the proto-Caribbean seaway occurs from 150 Ma through symmetrical seafloor spreading between the diverging North American and South America/African plates. For ease of reference, we will refer to this stage as the "proto-Caribbean and central Atlantic" opening. Break-up between the South American and African plates starts around 100 Ma with northward propagation from the south Atlantic. We refer to this second stage of seafloor spreading as "equatorial Atlantic" opening.

Most of the proto-Caribbean oceanic lithosphere has been subducted, but there remains a small segment in the south of the study area. The rifted oceanic lithosphere boundary between it and the equatorial Atlantic is visible in satellite gravity to the northwest of the Demerara Rise where it clearly acts as the termination point for a number of small fracture zones south of the Doldrums Fracture Zone (red ellipse, Extended Data Fig. 1b).

We first compared major Atlantic fracture zones in the region (15-20, Marathon, Mercurius, Vema and Doldrums) as detected in satellite gravity data to modelled flow lines according to the model of ref. [13] (Extended Data Fig. 1). Overall, the largest misfit between the two was ~50 km, and we assign this value to the positional uncertainty of these features (see below). The geometrical relationships between the two phases of seafloor spreading are particularly clear on the African side of the Atlantic, where the sediment cover is thin and the full sequence preserved (compared with the sedimented and partially subducted American side). The analysis showed that the southern two fracture

zones (Vema and Doldrums) have only just reached the Lesser Antilles trench, whereas the northern fracture zone (15-20) only grazes the Lesser Antilles subduction zone. None of these three fracture zones are therefore sources of hydration below the Lesser Antilles Arc.

Next, we refined the location of the proto-Caribbean / equatorial Atlantic Ocean boundary through time (Extended Data Fig. 2) based upon two observations. (1) The oldest section of the Marathon and Mercurius fracture zones can be well fitted by a flowline based entirely upon relative motion between North America and Africa. Therefore, this region must have lain entirely north of (or upon) the boundary between the central Atlantic and proto-Caribbean before opening of the equatorial Atlantic. 2) The major fracture zones to the south (Vema and Doldrums) can be well fitted by a flowline based entirely upon relative motion between South America and Africa. In this case, the far western extent of these fracture zones (which is constrained by symmetry with the clearly observable extent of fracture zones on the African side) must mark the edge of the proto-Caribbean oceanic crust in order for the Demerara Rise to close back against the African continental margin before initiation of equatorial Atlantic spreading (Extended Data Fig. 2a). Finally, the proto-Caribbean spreading ridge was placed mid-way between the separating North and South America plates, with a minimum number of transform faults inserted to satisfy the continental plate geometries.

Using this updated geometry for the proto-Caribbean/equatorial Atlantic boundary, and our computed flowlines for the Marathon, Mercurius and unnamed proto-Caribbean fracture zones, we model the subduction of these incoming plate features beneath the Caribbean plate from 50 Ma through to the present. Convergence azimuths and velocities between the Caribbean plate and the Atlantic are extracted directly from the model of ref. [13].

Projecting tectonic features onto the slab. To track the features properly once they enter the subduction zone and the slab begins to dip, it is necessary to adjust their horizontal velocities. To do this, we use three different assumptions for how the slab deforms as it enters the subduction zone. One endmember is the 'kinematic' approach outlined in ref. [49]. whereby features are assumed to follow streamlines over the surface of a slab with a fixed geometry, that is, minimal to no plate stretching during subduction. We use the slab geometry of ref. [30]. determined using local seismicity, and ref. [50], which is based on teleseismic tomography, for the regions that this first model does not cover. We also assume that the slab geometry remains fixed relative to the Caribbean plate for the modelled time period. In the other endmember, the slab is assumed to maintain its horizontal velocity and acquire an additional vertical sinking velocity, which would imply some amount of plate stretching. For the plate motions of the region, the first approach places incoming plate features further south than the second. We run a third, 'best estimate' model that is intermediate between the two.

Dehydration modelling. As incoming plate features move into the subduction zone, they dehydrate. Major pulses of subducting-plate dehydration occur[9] below the forearc and at sub-arc depths. Forearc dehydration includes the expulsion of pore fluids and the first breakdown of hydrous phases in the oceanic crust, while the sub-arc pulse starts with the blueschist transition that initiates directly below the maximum decoupling depth, below which the cool subducting plate first becomes coupled to the hot convecting mantle wedge. Following ref. [1] in computing phase stability fields, and using the kinematic thermal model set-up of ref. [51] to compute a thermal structure for the geometry and velocity of the Antilles slab, we predict that the first pulse of dehydration extends down to about 40 km depth, and the sub-arc pulse peaks at a depth up to 100–120 km (based on preliminary tomographic models by ref. [52].). In a similar model for the Greek subduction zone (which is similarly slow and old as the Antilles), the main dehydration depth intervals agree with regions of high V_p/V_s (that is, ratio of P wave to S wave velocity) above the slab, as expected from fluid release[53]. Motivated by these thermal models, sub-arc observations (number of Benioff zone earthquakes) and observations at the volcanic arc itself (boron isotopic signature, present-day volcanic output and crustal thickness) are compared at a dehydration depth of 100 km, which matches the average sub-arc slab depth. Comparisons with observations that reflect conditions beneath the forearc (forearc V_s and *b*-value anomalies) are done at a dehydration depth of 40 km.

For this study, our interest is in lateral variations in water input. We assume that the fracture zones and Atlantic-Proto-Caribbean boundary are all sources of excess slab hydration, that is, where the slab incorporates significantly larger quantities of water, mainly in the form of serpentinite, than in the plate away from the fracture zones, based on observations of similar structures offshore central America[54]. In the modelling, we apply the same Gaussian excess hydration profile with a width of 15 km to all these features (that is, in addition to the uniform background). This width is informed by the lateral extent of the V_p/V_s anomaly observed underneath the Marathon fracture zone on the incoming plate[56]. To put an order-of-magnitude estimate on the absolute values for the rate of excess hydration along the arc due to the subduction of each feature, we assume that the region of anomalous V_p/V_s corresponds to 50% serpentinized mantle lithosphere, and that half of this additional water is released under the forearc and half under the arc. We only model the along-strike variations in excess dehydration (that is, we set background hydration to zero).

We ultimately use the models to calculate the relative rate of hydration along the arc over the past 2 Myr for meaningful comparison with features that should depend on the present-day/recent dehydration below the arc and forearc, and over the past 25 Myr (the age of the current arc) for meaningful comparison with features that should depend on the total amount of water supplied to the arc (that is, the crustal thickness). The results of these calculations are presented in Extended Data Fig. 4 for a 'best estimate' calculation which uses the 'halfway' approach to slab deformation; a 'southern bound' calculation, which uses the stretched-slab endmember plus a 50-km shift to the south (the maximum misfit between our modelled fracture zones and the actual fracture zones on the African side of the Atlantic); and a 'northern bound' model, which uses the 'minimal-stretching' approach[49] plus a 50-km shift to the north.

Key results. If we take the best estimate model, we predict that the dehydration peak due to the Marathon and Mercurius fracture zones and the Proto-Caribbean/equatorial Atlantic plate boundary lies currently underneath Dominica (solid red line). In the main article, we demonstrate that this corresponds well with the peak in $\delta^{11}B$, sub-arc Wadati–Benioff earthquakes and volcanic output. We also predict that, if these three features are dehydrating underneath the forearc, then they would currently be doing so trenchwards of Martinique (dashed yellow line). This corresponds well with anomalies in V_s at a depth of around 50 km and the *b*-values for earthquakes in the forearc/plate-interface region. Looking at the full history of the arc (0–25 Ma; dotted blue line), there is a broad peak between Dominica and St Kitts and Nevis; the northern part of the arc. This higher rate of fluid flux in the north of the arc throughout the lifetime of the current arc may have resulted in a higher long-term magmatic output and therefore a thicker crust[7] if flux melting occurred. However, we cannot constrain the relative contribution of flux melting versus decompression melting. There are also peaks in the present-day dehydration rate and long-term dehydration rate in the far south of the arc between Grenada and St Vincent. These are due to the subduction of the unnamed proto-Caribbean fracture zone, the exact position of which is more speculative than for the Atlantic features. However, such features on the proto-Caribbean plate could potentially be responsible for the $\delta^{11}B$ anomaly observed at St Vincent.

Data availability

All geochemical data generated during this study are included in this published article (and its supplementary information files) and can be accessed in the EarthChem repository (https://doi.org/10.26022/IEDA/111527). Compiled geochemical data are freely available from the GEOROC database (http://georoc.mpch-mainz.gwdg.de/georoc/). Shear velocity model data can be accessed at https://doi.org/10.5258/SOTON/D1306. All broadband OBS data collected by the VoiLA project will become freely available through the IRIS Data Management Center via their data request tools, at the end of the project (April 2021).

Code availability

For plate-tectonic reconstructions we used the GPlates software, which is freely available at https://www.gplates.org/ with the plate model at https://www.earthbyte.org/global-plate-models.

36. Humphreys, M. C. S., Kearns, S. L. & Blundy, J. D. SIMS investigation of electron-beam damage to hydrous, rhyolitic glasses: implications for melt inclusion analysis. *Am. Mineral.* **91**, 667–679 (2006).
37. Boschi, C. et al. Serpentinization of mantle peridotites along an uplifted lithospheric section, Mid Atlantic Ridge at 11° N. *Lithos* **178**, 3–23 (2013).
38. Boschi, C., Dini, A., Früh-Green, G. L. & Kelley, D. S. Isotopic and element exchange during serpentinization and metasomatism at the Atlantis Massif (MAR 30° N): insights from B and Sr isotope data. *Geochim. Cosmochim. Acta* **72**, 1801–1823 (2008).
39. Spivack, A. J. & Edmond, J. M. Boron isotope exchange between seawater and the oceanic crust. *Geochim. Cosmochim. Acta* **51**, 1033–1043 (1987).
40. Marschall, H. R. et al. The boron and lithium isotopic composition of mid-ocean ridge basalts and the mantle. *Geochim. Cosmochim. Acta* **207**, 102–138 (2017).
41. Workman, R. K. & Hart, S. R. Major and trace element composition of the depleted MORB mantle (DMM). *Earth Planet. Sci. Lett.* **231**, 53–72 (2005).
42. Collier, J. S. *VoiLA—Volatile Recycling in the Lesser Antilles Arc: RRS* James Cook *Cruise Report JC133*, 79, https://www.bodc.ac.uk/resources/inventories/cruise_inventory/reports/jc133.pdf (2015).
43. Collier, J. S. *VOILA—Volatile Recycling in the Lesser Antilles Arc: RRS* James Cook *Cruise Report JC149*, 161, https://www.bodc.ac.uk/resources/inventories/cruise_inventory/reports/jc149.pdf (2017).
44. Landisman, M., Dziewonski, A. & Satô, Y. Recent improvements in the analysis of surface wave observations. *Geophys. J. Int.* **17**, 369–403 (1969).
45. Levshin, A. L. & Ritzwoller, M. H. in *Monitoring the Comprehensive Nuclear-Test-Ban Treaty: Surface Waves* (eds Levshin, A. L. & Ritzwoller, M. H.) 1531–1545 (Birkhäuser Basel, 2001).
46. Harmon, N. & Rychert, C. A. Joint inversion of teleseismic and ambient noise Rayleigh waves for phase velocity maps, an application to Iceland: noise-teleseismic phase velocity maps. *J. Geophys. Res. Solid Earth* **121**, 5966–5987 (2016).
47. Forsyth, D. W. & Li, A. in *Seismic Earth: Array Analysis of Broadband Seismograms* (eds Levander, A. & Nolet, G.) 81–97 (AGU, 2005).
48. Yang, Y. & Forsyth, D. W. Regional tomographic inversion of the amplitude and phase of Rayleigh waves with 2-D sensitivity kernels. *Geophys. J. Int.* **166**, 1148–1160 (2006).
49. Harmon, N. et al. Mapping geologic features onto subducted slabs. *Geophys. J. Int.* **219**, 725–733 (2019).
50. Braszus, B. *3D Teleseismic Travel Time Tomography along the Lesser Antilles Subduction Zone*. MSc thesis, Karlsruhe Institute of Technology (2019).
51. Perrin, A. et al. Reconciling mantle wedge thermal structure with arc lava thermobarometric determinations in oceanic subduction zones. *Geochem. Geophys. Geosyst.* **17**, 4105–4127 (2016).
52. Hicks, S. P. et al. Evidence for an anomalously large cold mantle wedge corner of the Caribbean Plate in the Lesser Antilles subduction zone. In *AGU Fall Meeting 2019* T23A-08 (AGU, 2019).
53. Halpaap, F. et al. Earthquakes track subduction fluids from slab source to mantle wedge sink. *Sci. Adv.* **5**, eaav7369 (2019).
54. Van Avendonk, H. J. A., Holbrook, W. S., Lizarralde, D. & Denyer, P. Structure and serpentinization of the subducting Cocos plate offshore Nicaragua and Costa Rica. *Geochem. Geophys. Geosystems* **12**, Q06009 (2011).
55. Sandwell, D. T., Müller, R. D., Smith, W. H. F., Garcia, E. & Francis, R. New global marine gravity model from CryoSat-2 and Jason-1 reveals buried tectonic structure. *Science* **346**, 65–67 (2014).
56. Davy, R. G., Collier, J. S., Henstock, T. J. & the VoiLA consortium. Wide-angle seismic imaging of two modes of crustal accretion in mature Atlantic Ocean crust. *J. Geophys. Res. Solid Earth* https://doi.org/10.1029/2019JB019100 (2020).

Acknowledgements We thank our project partners R. Robertson, J. Latchman, S. Tait and F. Krüger for support and discussion over the course of this project. We thank C. J. de Hoog for assistance with SIMS analysis at the EIMF, S. Kearns for help with EPMA analysis, the German Instrument Pool for Amphibian Seismology (DEPAS), hosted by the Alfred Wegener Institute Bremerhaven, for providing the ocean-bottom and temporary island seismometers, and the Scripps Institution of Oceanography (OBSIP) for providing additional ocean-bottom seismometers. This research was funded by the VoiLA NERC consortium grant (NE/K010824/1). SIMS analysis was funded by EIMF proposals IMF619/0517 and IMF653/0518.

Author contributions All authors discussed the results and implications of the work and commented on the manuscript at all stages. G.F.C., C.G.M., J.D.B. and A.A.I carried out geochemical analysis and interpretation. G.F.C., S.G., C.G.M., J.D.B. and J.C. drafted the manuscript. N.H. and C.R. produced the shear-wave velocity model. B.M. made the dehydration model. L.B. and S.P.H. compiled local seismicity data. D.S. mapped *b*-values. R.W.A and J.C. produced the tectonic reconstruction and associated figures. C.G.M., S.G., J.D.B., J.C., A.R., N.H., C.R., J.P.D., T.J.H., J.v.H., J.J.W. and M.W. designed the original VoiLA experiment.

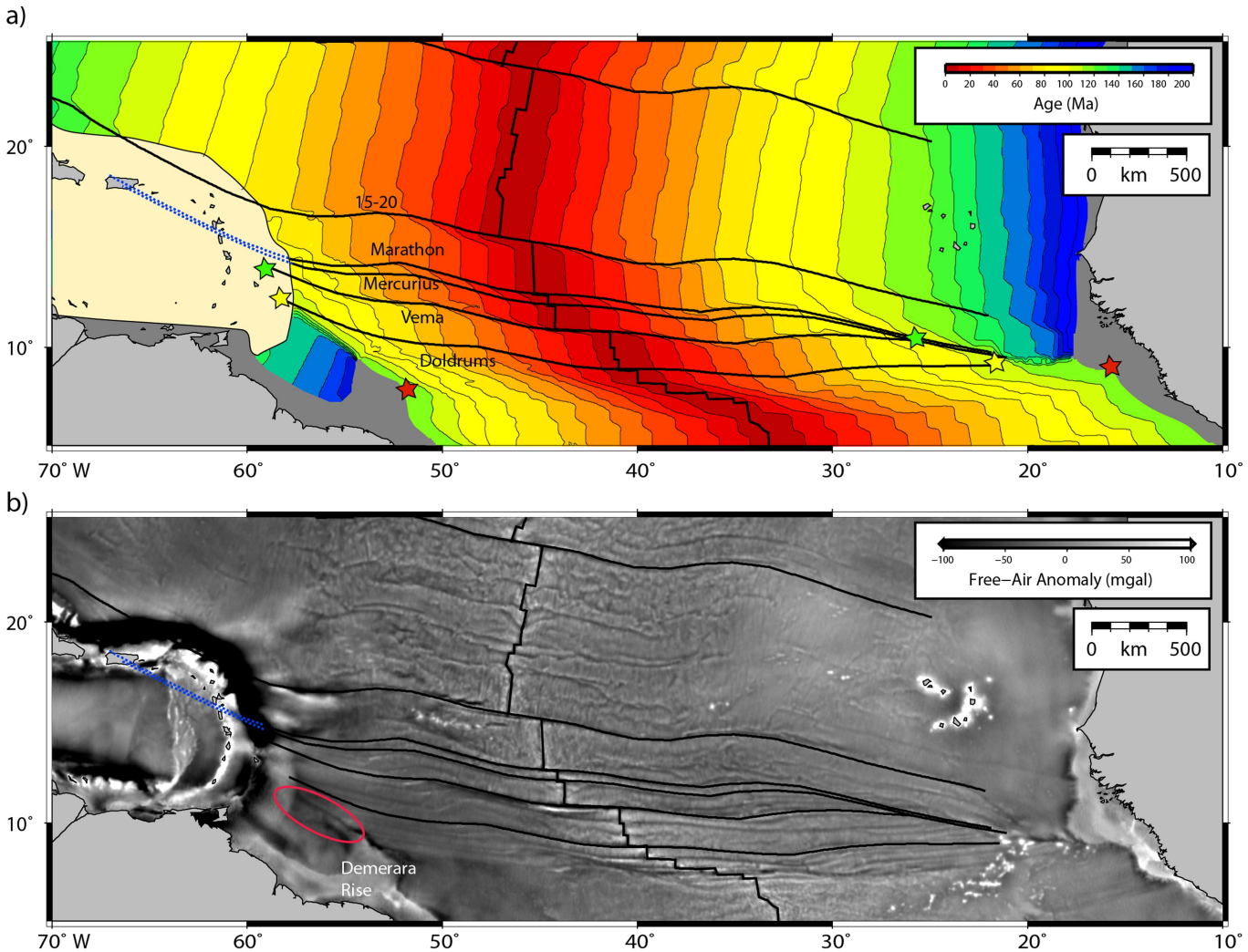
Competing interests The authors declare no competing interests.

Additional information

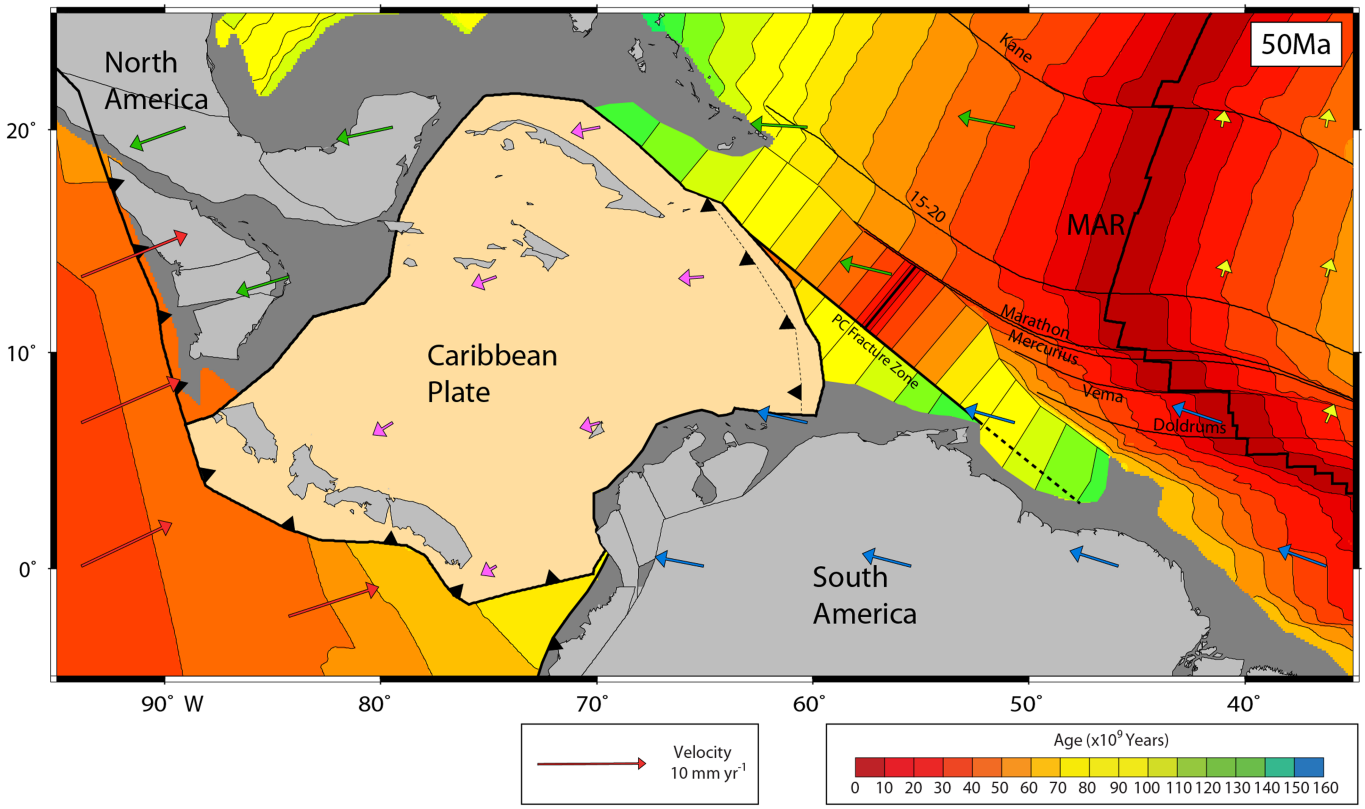
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2407-5.

Correspondence and requests for materials should be addressed to G.F.C.

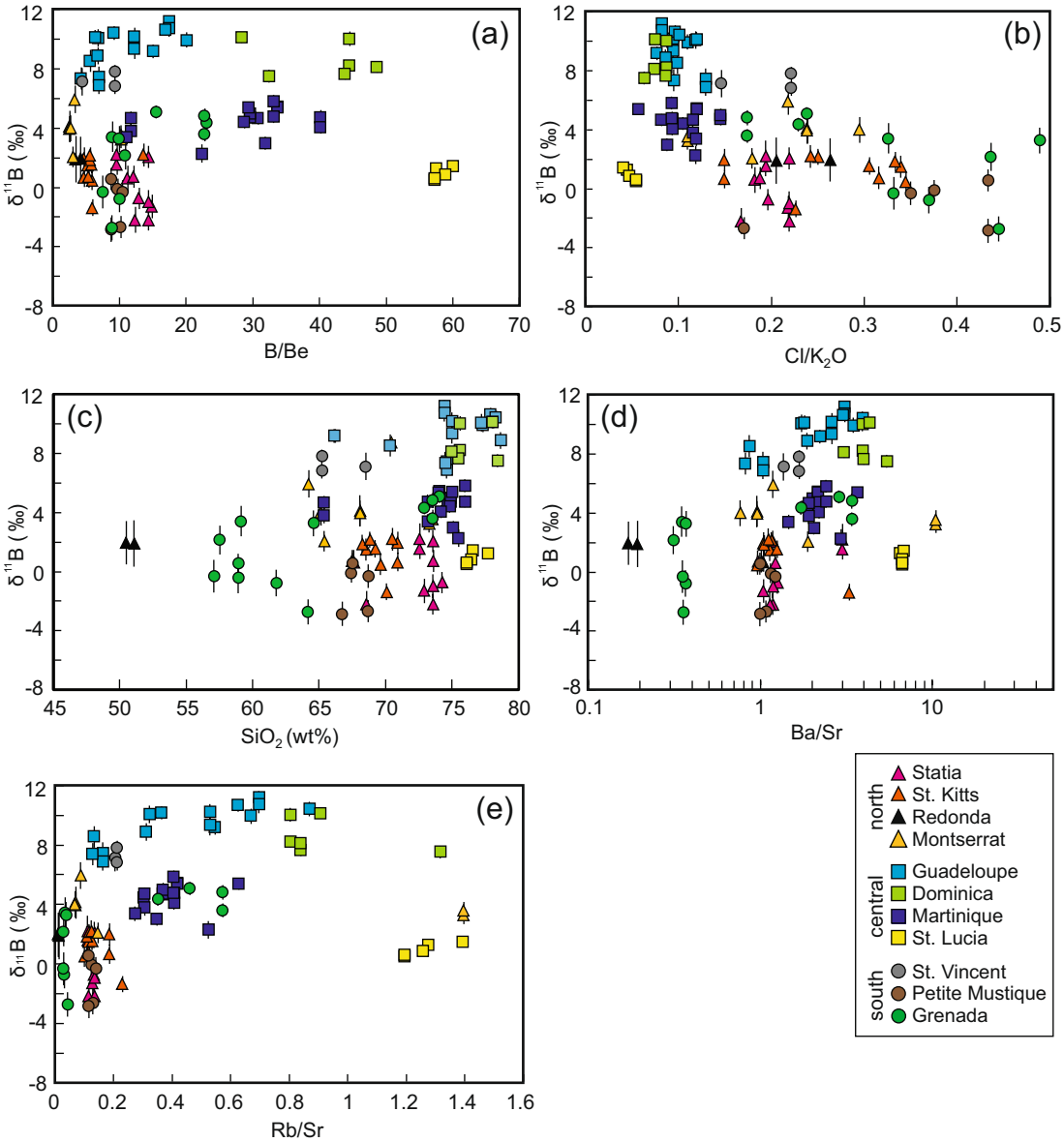
Peer review information *Nature* thanks William Leeman and Othmar Müntener for their contribution to the peer review of this work.



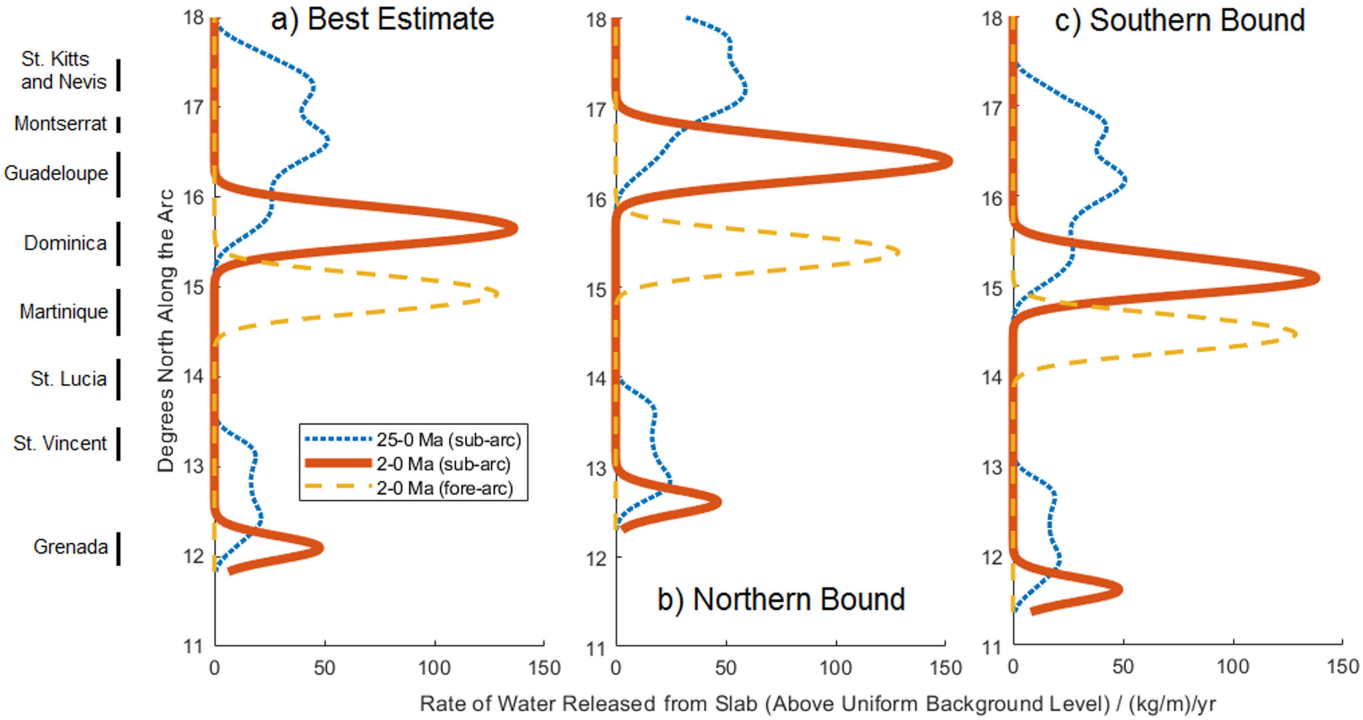
Extended Data Fig. 1 | Modelled fracture zones. a, Modelled fracture zones in the central Atlantic overlaid on an oceanic crust age grid from ref. [13]. Coloured stars denote conjugate points associated with opening of the equatorial Atlantic at either end of the Vema (green stars) and Doldrums (yellow stars) fracture zones, and between the Demerara Rise and African continental margin (red stars). **b**, Modelled fracture zones overlaid on satellite free-air gravity[55]. Red ellipse marks the location of the proto-Caribbean/Atlantic boundary.



Extended Data Fig. 2 | Modified plate reconstruction. Snapshot of modified plate reconstruction at 50 Ma (ref. [13]). Velocity vectors (coloured by plate) shown are relative to the mantle reference frame. The figure shows the four sources of dehydration from the subducted slab over the past 25 Myr considered here: (i) Marathon FZ; (ii) Mercurius FZ; (iii) proto-Caribbean/equatorial Atlantic boundary and (iv) unnamed FZ formed during proto-Caribbean opening, labelled PC Fracture Zone. MAR, Mid-Atlantic Ridge.



Extended Data Fig. 3 | Melt inclusion $\delta^{11}B$. a–e, All values of melt inclusion $\delta^{11}B$ measured in this study are shown versus indicators of fluid composition (**a**, **b**) and differentiation (**c**–**e**). No clear observable trends are shown between islands, indicating that these differences are largely controlled by the mantle source.



Extended Data Fig. 4 | Excess dehydration. The average rate of excess dehydration (above a uniform background), resulting from the subduction of fracture zones and the proto-Caribbean/Atlantic plate boundary, along the arc from 11° N to 18° N over the past 2 Myr (red solid curve) and 25 Myr (blue dotted curve), and below the forearc over the past 2 Myr (dashed yellow line). The pattern of relative distribution of dehydration is robust, constrained by the history of fracture-zone/plate-boundary subduction, but the absolute values of the dehydration rates should be treated with caution, as they depend strongly on the simple model assumptions of the level of hydration and relative strength of fore- and sub-arc dehydration. **a**, Best estimate; **b**, northern bound endmember; **c**, southern bound (see text for details).

Extended Data Table 1 | $\delta^{11}B$ values, B concentrations and Nb/B of sources of fluids used in the mixing model (Fig. 3).

	DM Mantle	Serpentinite fluid	AOC & Sed (120km)	Residual slab melt
$\delta^{11}B$	-7.1	12	-3.9	-12.8
B (ppm)	0.077	289	170	16.8
Nb/B	1.93	0.00	0.01	0.01

Article

Monumental architecture at Aguada Fénix and the rise of Maya civilization

https://doi.org/10.1038/s41586-020-2343-4

Received: 26 November 2019

Accepted: 30 April 2020

Published online: 3 June 2020

Check for updates

Takeshi Inomata[1]✉, Daniela Triadan[1], Verónica A. Vázquez López[2], Juan Carlos Fernandez-Diaz[3], Takayuki Omori[4], María Belén Méndez Bauer[5], Melina García Hernández[6], Timothy Beach[7], Clarissa Cagnato[8], Kazuo Aoyama[9] & Hiroo Nasu[10]

Archaeologists have traditionally thought that the development of Maya civilization was gradual, assuming that small villages began to emerge during the Middle Preclassic period (1000–350 BC; dates are calibrated throughout) along with the use of ceramics and the adoption of sedentism[1]. Recent finds of early ceremonial complexes are beginning to challenge this model. Here we describe an airborne lidar survey and excavations of the previously unknown site of Aguada Fénix (Tabasco, Mexico) with an artificial plateau, which measures 1,400 m in length and 10 to 15 m in height and has 9 causeways radiating out from it. We dated this construction to between 1000 and 800 BC using a Bayesian analysis of radiocarbon dates. To our knowledge, this is the oldest monumental construction ever found in the Maya area and the largest in the entire pre-Hispanic history of the region. Although the site exhibits some similarities to the earlier Olmec centre of San Lorenzo, the community of Aguada Fénix probably did not have marked social inequality comparable to that of San Lorenzo. Aguada Fénix and other ceremonial complexes of the same period suggest the importance of communal work in the initial development of Maya civilization.

The period around 1200–1000 BC was a critical time of social change in the Maya lowlands. Prior to this period, the inhabitants of this area did not use ceramics and probably maintained mobile ways of life by combining hunting, gathering and fishing with the cultivation of maize and other crops[2]. They began to adopt ceramics and greater degrees of sedentism at the beginning of the Middle Preclassic period, and researchers have long thought that ceremonial centres with large pyramids did not develop until late in the Middle Preclassic period, or in the Late Preclassic and Terminal Preclassic periods (hereafter, Late–Terminal Preclassic) (350 BC–AD 250). However, the discovery of a formal ceremonial complex and an artificial plateau at Ceibal dating to 950 BC suggests that substantial ceremonial centres developed in the Maya lowlands earlier than was previously thought[3,4]. Here, the term artificial plateau refers to horizontal buildings larger than 200 × 200 m, as distinguished from smaller supporting platforms. A few centuries later, other centres in the Maya lowlands—such as Cival, Komchen, Nakbe, Yaxnohcah and Xocnaceh—also built artificial plateaus or large platforms[5–9]. Our research in Tabasco (Mexico) has revealed an even older and larger ceremonial centre, Aguada Fénix (Extended Data Fig. 1).

We began the Middle Usumacinta Archaeological Project in the area along the Usumacinta and San Pedro Rivers in Tabasco in 2017 (Fig. 1). Despite previous investigations[10,11] in the area, the Preclassic period of this region was poorly understood. We thought that this area, located at the western periphery of the Maya lowlands, might hold the key to understanding the relationship between the Olmec civilization and Maya society. The Olmec centre of San Lorenzo—which reached its heyday between 1400 and 1150 BC—is characterized by an enormous artificial plateau and colossal sculptures of stone heads, but does not have pyramids[12,13]. During the Middle Preclassic period (possibly after 800 BC), La Venta became a dominant Olmec centre, containing a large pyramid and mounds[14–16]. Archaeologists have long debated whether the inhabitants of the Maya lowlands inherited the legacy of San Lorenzo, and whether they received direct influence from La Venta[17,18].

Survey and excavation

A high-resolution lidar survey conducted by the National Center for Airborne Laser Mapping (NCALM) and a low-resolution lidar survey by the Instituto Nacional de Estadística y Geografía (INEGI) in our study area revealed 21 ceremonial centres in a standardized spatial configuration, which we call the Middle Formative Usumacinta (MFU) pattern. The MFU pattern is characterized by a rectangular shape defined by rows of low mounds, oriented roughly north–south (Fig. 2). At the centre of each MFU complex is a so-called E-group assemblage, which consists of a round or square western mound and an elongated eastern platform. Many other sites in the Maya lowlands that date to the Middle Preclassic

[1]School of Anthropology, University of Arizona, Tucson, AZ, USA. [2]Department of Anthropology and Archaeology, University of Calgary, Alberta, Canada. [3]National Center for Airborne Laser Mapping (NCALM), University of Houston, Houston, TX, USA. [4]University Museum, University of Tokyo, Tokyo, Japan. [5]Estudios Mesoamericanos, Universidad Nacional Autónoma de México, Mexico City, Mexico. [6]Middle Usumacinta Archaeological Project, Balancán, Mexico. [7]Department of Geography and the Environment, University of Texas, Austin, TX, USA. [8]UMR 8096, Archéologie des Amériques, Nanterre, France. [9]Faculty of Humanities, Ibaraki University, Mito, Japan. [10]Faculty of Biosphere-Geosphere Science, Okayama University of Science, Okayama, Japan. ✉e-mail: inomata@arizona.edu

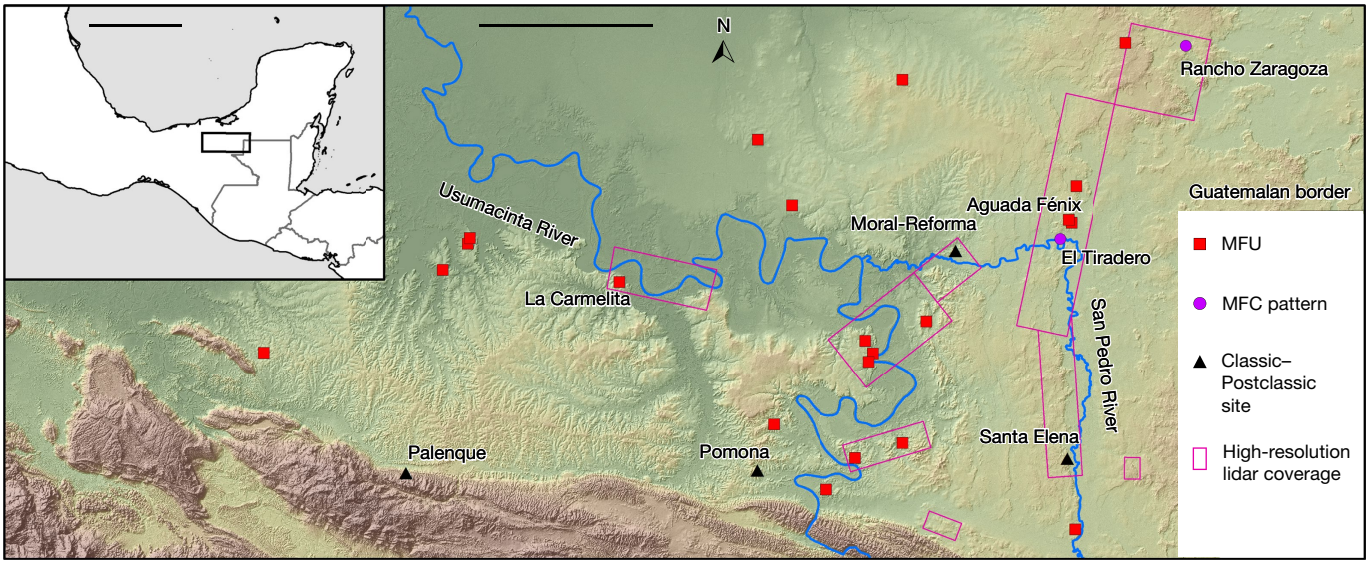


Fig. 1 | Map of the Middle Usumacinta region. The low-resolution INEGI lidar covers the entire region. MFC, Middle Formative Chiapas. Map topographic data are from the INEGI lidar survey (www.inegi.org.mx). Scale bars, 25 km (main panel), 400 km (inset).

and Late–Terminal Preclassic periods have E-group assemblages, but no rectangular site plans are found to the east of our study area[19]. We also found smaller versions of the MFU complex—measuring less than 400 m in length—that we call 'minor MFU' complexes. Moreover, there are roughly rectangular complexes that exhibit less formal shapes, without a clear E-group assemblage.

The MFU pattern is probably related to what has previously been referred to as the Middle Formative Chiapas pattern, which is found at sites of the Middle Preclassic period (including La Venta, centres in the Grijalva River basin, Tzutzuculi and Ceibal)[17,20]. The Middle Formative Chiapas pattern consists of an E-group assemblage and large platforms that are arranged along a north–south axis, but lacks the delineated rectangular shape of the MFU pattern. Middle Formative Chiapas complexes appear to have been built between 1000 and 350 BC. Excavations at the Middle Formative Chiapas centres of La Venta, San Isidro, Chiapa de Corzo and Ceibal have unearthed a series of caches with greenstone axes: these communities probably shared similar ritual concepts and practices[14,21–23].

The largest of the MFU sites in our study region is Aguada Fénix (Fig. 1). The high-resolution lidar shows that the main plateau of this site has a rectangular form, measuring 1,413 m from north to south and 399 m from east to west, and that its edges were lined with low platforms. Square wings attached to the eastern and western sides of this plateau give it a narrow cruciform-like overall shape (Fig. 2). The large southwest platform may have been added later to this original form. Unlike other MFU sites (which do not have substantial build-ups of plaza areas), this construction rises 10 to 15 m above the surrounding ground surface. This site was not known before our research, probably because a horizontal construction on this scale is difficult to recognize from the ground level. A large E-group assemblage, with the eastern platform measuring 401 m in length, occupies the centre of the formation. The plateau is surrounded by one MFU complex, five minor MFU complexes, multiple rectangular complexes and artificial reservoirs, as well as by wetlands on the east. In addition, nine causeways extend from the plateau. The northern and southern causeways are connected to the plateau by large ramps. The northwest causeway is the longest of all and extends 6.3 km, connecting multiple complexes along the way (Extended Data Fig. 2). The west plateau is another large construction, measuring 390 × 270 m horizontally and 15–18 m in height; it stands 1.7 km to the west of the main plateau.

Our excavation results indicate that the main plateau was raised multiple times with clay and earth fills, and reached a size close to the current one around 800 BC. In the 7.5-m-deep operation NR3A (for definitions of excavation designations, see 'Excavation' in Methods), we uncovered a dense deposit of ceramics, bones and shells covering bedrock, which appears to predate the construction of the plateau (Extended Data Figs. 3, 4). The plateau construction events included two episodes, in which clays and other soils of various colour were placed in multiple layers, each layer forming checkerboard-like horizontal patterns (Extended Data Fig. 5). The presence of similar—albeit thinner—fills in operations NR5A, NR7A and NR9A indicates that the builders placed elaborate fills of multiple colours over a large part of the plateau, which they covered with a floor at the end of each construction event.

The results from operation NR7A showed that this edge platform was also constructed mostly with earthen fills during the Middle Preclassic period. Nevertheless, four structures located directly west of the E group have walls made of roughly shaped megalithic blocks (Extended Data Fig. 6). Operation NR8A revealed blocks measuring up to 3.0 × 1.0 × 0.7 m. Through excavations in two of the causeways (operations NR4A and NR6A), we also confirmed that these wide streets were built during the Middle Preclassic period, with fill thicknesses of around 2.6 m.

Radiocarbon dates

We obtained 69 radiocarbon dates, which we analysed using Bayesian statistics (Extended Data Fig. 7, Supplementary Methods, Supplementary Data, Supplementary Table 1). Charcoal samples from the earliest deposits in operations NR3A and NR7A at Aguada Fénix yielded dates of around 1250–1150 BC and 1150–1050 BC, respectively. These data indicate that the people of this region had begun to use ceramics by 1200 BC, one to two centuries earlier than those of Ceibal, Tikal, Cahal Pech, Cuello and other Maya communities[2]. Plateau construction began by 1000 BC if not earlier, slightly before the initial construction of the ceremonial complex at Ceibal. However, construction activity at Aguada Fénix ceased soon after 800 BC. Carbon samples from two of the causeways yielded radiocarbon dates of 950–800 BC. In addition, samples taken from test excavations in areas around the plateau, where residences may have existed, returned dates of 1000–750 BC. At the MFU site of La Carmelita, carbon samples from the lowest layer yielded dates of around 900 BC, and samples from the upper layers gave dates of around 750 BC (Extended Data Fig. 8). We suspect that other

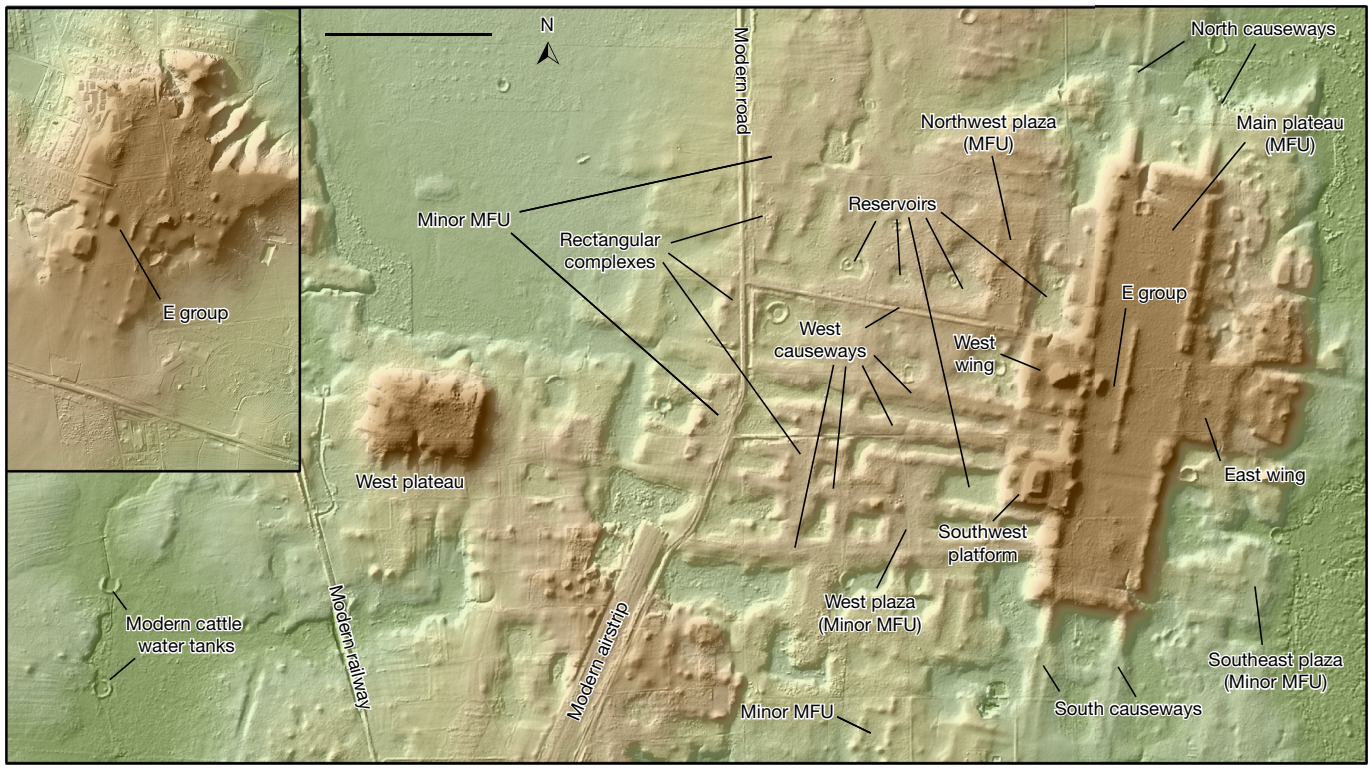


Fig. 2 | High-resolution lidar images of Aguada Fénix and La Carmelita. Main panel, Aguada Fénix; inset, La Carmelita. Scale bar, 500 m (both images are on the same scale).

MFU sites in the region were also built during the period between 1000 and 750 BC. Aguada Fénix and other MFU sites appear to have been abandoned by 750 BC. Small groups returned to Aguada Fénix during the Late–Terminal Preclassic and Late Classic periods.

Volume estimates

In addition to the excavations, we conducted auger tests in the main and west plateaus at Aguada Fénix to estimate their construction volumes. The results suggest that the builders constructed the main plateau over a natural rise of bedrock (Extended Data Fig. 9a, Supplementary Table 2). On the basis of the reconstructed bedrock surface and lidar data, we estimate the fill volume for the Middle Preclassic portion of the main plateau at 3,200,000–4,300,000 m^3. We calculate that this Middle Preclassic construction required 10,000,000–13,000,000 person-days (Extended Data Fig. 9b).

The volume of the main plateau surpasses that of the La Danta complex at the Late–Terminal Preclassic centre of El Mirador, the largest construction previously known in the Maya lowlands[7] (Extended Data Fig. 9c). Pyramids built during the Classic period in the Maya lowlands are substantially smaller[24]. In other words, the main plateau of Aguada Fénix is the largest construction in the pre-Hispanic Maya area. The volume of the plateau at San Lorenzo is larger but after the decline of this Olmec centre, Aguada Fénix represented the largest construction effort during the Middle Preclassic and Late–Terminal Preclassic periods in Mesoamerica[13]. It is noteworthy that this enormous construction at Aguada Fénix was built in a short span, of roughly 200 years.

Discussion

Artificial plateaus may be characterized as horizontal monumentality, which contrasts with the vertical dimensions of pyramids. The construction of the plateaus at Aguada Fénix most probably followed the tradition established at San Lorenzo. The builders combined this legacy of the previous era with elements that emerged after the decline of San Lorenzo, including standardized site plans, the E-group assemblage and other pyramidal constructions. These innovations probably occurred through intensive interregional interaction. The Pacific coast may have been an important area for the development of pyramidal structures[25,26]. Aguada Fénix and other MFU complexes shared standardized spatial configurations and the E-group assemblage with the Middle Formative Chiapas centres in the Grijalva River region. A greenstone axe cache found in the E-group plaza of Aguada Fénix indicates that its inhabitants also practiced rituals similar to those of La Venta, the Grijalva River region and Ceibal (Extended Data Fig. 10). Aguada Fénix appears to have had a central role in this dynamic process of social and cultural innovation between 1100 and 800 BC.

Despite their architectural and ritual commonality, the political and cultural settings of these regions were diverse. The ceramics found at Aguada Fénix resemble the Real ceramics from Ceibal and are markedly different from those of the La Venta or the Grijalva River region. Although the ceramics do not necessarily indicate that the builders of Aguada Fénix were speakers of a Mayan language, they appear to have had closer cultural affinities with the Maya lowlands than with the Olmec area. This interpretation is bolstered by the observation that all analysed obsidian pieces from our study area originated from El Chayal and other Guatemalan sources (Supplementary Table 3). This finding contrasts with the pattern at San Lorenzo, where a substantial portion of obsidian was imported from Mexican sources[27]. It is also likely that social inequality at Aguada Fénix was not as pronounced as at San Lorenzo and La Venta. Unlike those Olmec centres, Aguada Fénix does not exhibit clear indicators of marked social inequality, such as sculptures representing high-status individuals. The only stone sculpture found so far at Aguada Fénix depicts an animal (Extended Data Fig. 10). If these interpretations are correct, they imply that the Gulf Coast Olmec region was not the only centre of cultural development and that innovations did not always emanate from the most hierarchical polities.

An important factor for the emergence of Aguada Fénix and related sites may have been the transition from a mobile lifeway to sedentism, stimulated by a heavier reliance on maize agriculture[2,28–31]. The scarcity of residential platforms around many of the MFU sites suggests that a substantial portion of the inhabitants of the Middle Usumacinta region maintained a degree of residential mobility. At the same time, results from the analysis of starch grains found on grinding stones are consistent with the assumption that the use of maize was common during the period of plateau construction (Supplementary Table 4). Under rapidly changing social conditions, many inhabitants of the region may have actively participated in the transformation of the lived landscape to create new places of gathering without coercion from powerful elites. Although the tradition of horizontal monumentality was first established at the hierarchical polity of San Lorenzo, the inclusive forms of plateaus may have been appealing to communities without marked social inequality. With the development of more hierarchical organization, later sites—including La Venta, Takalik Abaj, Nakbe and Tikal—emphasized tall pyramids, access to which was possibly restricted to a privileged few.

Aguada Fénix may be analogous to early ceremonial constructions that emerged during pre-agricultural or incipient agricultural periods in other parts of the world, including the Near East, the Andes and the American Southeast[32–36]. However, Aguada Fénix is different from these examples in that Mesoamerican groups had domesticated maize and other crops several millennia before the rise of Aguada Fénix[37]. These observations urge us to explore the diverse processes that existed in the construction of monumental structures in societies with limited social inequality.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2343-4.

1. Adams, R. E. W. *The Origins of Maya Civilization* (Univ. New Mexico Press, 1977).
2. Lohse, J. C. Archaic origins of the lowland Maya. *Lat. Am. Antiq.* **21**, 312–352 (2010).
3. Inomata, T., Triadan, D., Aoyama, K., Castillo, V. & Yonenobu, H. Early ceremonial constructions at Ceibal, Guatemala, and the origins of lowland Maya civilization. *Science* **340**, 467–471 (2013).
4. Inomata, T., Triadan, D., Pinzón, F. & Aoyama, K. Artificial plateau construction during the Preclassic period at the Maya site of Ceibal, Guatemala. *PLoS ONE* **14**, e0221943 (2019).
5. Estrada-Belli, F. in *Early New World Monumentality* (eds Burger, R. L. & Rosenswig, R. M.) 198–230 (Univ. Press Florida, 2012).
6. Andrews, E. W. V, Bey, G. J. III & Gunn, C. M. in *Pathways to Complexity: A View from the Maya Lowlands* (eds Brown, M. K. & Bey, G. J. III) 49–86 (Univ. Press Florida, 2018).
7. Hansen, R. D. & Suyuc, L. E. *Mirador* (FARES Guatemala, 2016).
8. Gallareta Negrón, T. in *Pathways to Complexity: A View from the Maya Lowlands* (eds Brown, M. K. & Bey, G. J. III) 276–291 (Univ. Press Florida, 2018).
9. Reese-Taylor, K. in *Maya E Groups: Calendars, Astronomy, and Urbanism in the Early Lowlands* (eds Freidel, D. A. et al.) 480–513 (Univ. Press Florida, 2017).
10. Rands, R. L. in *Origins of Maya Civilization* (ed. Adams, R. E. W.) 159–180 (Univ. New Mexico Press, 1977).
11. Ochoa, L. in *Antropología e Historia de los Mixe-Zoques y Mayas: Homenaje a Frans Blom* (eds Ochoa, L. & Lee, T. A.) 147–174 (Instituto de Investigaciones Filológicas, UNAM, 1983).
12. Coe, M. D. & Diehl, R. A. *In the Land of the Olmec* (Univ. Texas Press, 1980).
13. Cyphers, A. in *The Origins of Maya States* (eds Traxler, L. P. & Sharer, R. J.) 83–122 (Univ. Pennsylvania Museum of Archaeology and Anthropology, 2016).
14. Drucker, P., Heizer, R. F. & Squier, R. H. *Excavations at La Venta, Tabasco, 1955* (Smithsonian Institution, 1959).
15. González Lauck, R. B. in *The Place of Stone Monuments: Context, Use, and Meaning in Mesoamerica's Preclassic Tradition* (eds Guernsey, J. et al.) 177–205 (Dumbarton Oaks Research Library and Collection, 2010).
16. Clark, J. E. in *The Origins of Maya States* (eds Traxler, L. P. & Sharer, R. J.) 123–224 (Univ. Pennsylvania Museum of Archaeology and Anthropology, 2016).
17. Clark, J. E. & Hansen, R. D. in *Royal Courts of the Ancient Maya, Volume 2: Data and Case Studies* (eds Inomata, T. & Houston, S. D.) 1–45 (Westview Press, 2001).
18. Estrada-Belli, F. *The First Maya Civilization: Ritual and Power before the Classic Period* (Routledge, 2011).
19. Freidel, D. A., Chase, A. F., Dowd, A. S. & Murdock, J. *Maya E Groups: Calendars, Astronomy, and Urbanism in the Early Lowlands* (Univ. Press Florida, 2017).
20. Lowe, G. W. in *The Origins of Maya Civilization* (ed. Adams, R. E. W.) 197–248 (Univ. New Mexico Press, 1977).
21. Lowe, G. W. in *The Olmec and their Neighbors* (eds Coe, M. D. & Grove, D.) 231–256 (Dumbarton Oaks Research Library and Collection, 1981).
22. Bachand, B. R. & Lowe, L. S. in *Arqueología Reciente de Chiapas: Contribuciones del Encuentro Celebrado en el 60° Aniversario de la Fundación Arqueológica Nuevo Mundo* (eds Lowe, L. S. & Pye, M. E.) 45–68 (Brigham Young Univ., 2012).
23. Inomata, T. & Triadan, D. Middle Preclassic Caches from Ceibal, Guatemala. *Maya Archaeol.* **3**, 56–91 (2016).
24. Webster, D. & Kirker, J. Too many Maya, too few buildings: investigating construction potential at Copán, Honduras. *J. Anthropol. Res.* **51**, 363–387 (1995).
25. Hodgson, J. G., Clark, J. G. & Gallaga Murrieta, E. Ojo de Agua monument 3: a new Olmec-style sculpture from Ojo de Agua, Chiapas, Mexico. *Mexicon* **32**, 139–144 (2010).
26. Love, M. & Guernsey, J. in *Early Mesoamerican Social Transformations: Archaic and Formative Lifeways in the Soconusco Region* (ed. Lesure, R. G.) 170–188 (Univ. California Press, 2011).
27. Hirth, K., Cyphers, A., Cobean, R., De León, J. & Glascock, M. D. Early Olmec obsidian trade and economic organization at San Lorenzo. *J. Archaeol. Sci.* **40**, 2784–2798 (2013).
28. Blake, M., Clark, J. E., Voorhies, B., Love, M. W. & Chisholm, B. S. Prehistoric subsistence in the Soconusco region. *Curr. Anthropol.* **33**, 83–94 (1992).
29. Clark, J. E., Pye, M. E. & Gosser, D. C. in *Archaeology, Art, and Ethnogenesis in Mesoamerican Prehistory: Papers in Honor of Gareth W. Lowe* (eds Lowe, L. S. & Pye, M. E.) 23–42 (Brigham Young Univ., 2007).
30. Inomata, T. et al. Development of sedentary communities in the Maya lowlands: coexisting mobile groups and public ceremonies at Ceibal, Guatemala. *Proc. Natl Acad. Sci. USA* **112**, 4268–4273 (2015).
31. Rosenswig, R. M. in *Early Mesoamerican Social Transformations: Archaic and Formative Lifeways in the Soconusco Region* (ed. Lesure, R. G.) 242–271 (Univ. California Press, 2011).
32. Schmidt, K. Göbekli Tepe–the Stone Age sanctuaries: new results of ongoing excavations with a special focus on sculptures and high reliefs. *Documenta Praehistorica* **37**, 239–255 (2010).
33. Solis, R. S., Haas, J. & Creamer, W. Dating Caral, a preceramic site in the Supe Valley on the central coast of Peru. *Science* **292**, 723–726 (2001).
34. Saunders, J. W. et al. Watson Brake, a Middle Archaic mound complex in northeast Louisiana. *Am. Antiq.* **70**, 631–668 (2005).
35. Stanish, C. *The Evolution of Human Co-Operation: Ritual and Social Complexity in Stateless Societies* (Cambridge Univ. Press, 2017).
36. Burger, R. L. & Rosenswig, R. M. *Early New World Monumentality* (eds Burger, R. L. & Rosenswig, R. M.) (Univ. Press Florida, 2012).
37. Piperno, D. R., Ranere, A. J., Holst, I., Iriarte, J. & Dickau, R. Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico. *Proc. Natl Acad. Sci. USA* **106**, 5019–5024 (2009).

Methods

Lidar

Lidar data are now commonly used in archaeological investigations in southern Mesoamerica, as well as in other tropical regions of the world[38–51]. In our research, the high-resolution lidar data were obtained by NCALM. The NCALM crew collected lidar data for 109 km^2 on 6 May 2017. After the discovery of Aguada Fénix, we acquired additional lidar data for a nominal area of 745 km^2 between 9 June 2019 and 17 June 2019. The site of La Carmelita was surveyed in the 2017 NCALM campaign, and the entire extent of Aguada Fénix was covered by the NCALM high-resolution lidar data of 2019.

For both campaigns, the NCALM team used an Optech Titan lidar system, which is equipped with three channels of laser at wavelengths of 1,550, 1,064 and 532 nm[52,53]. The following parameters were used for the 2019 survey: a flying height of 650 m above ground level; a pulse repetition frequency of 150 kHz; a scan frequency of 25 Hz; and a scan angle of ±30°. This configuration produced swath widths of 750 m, which were laterally overlapped by 50%, with a flight line spacing of 345 m. Assessed over a 298.2-km^2 section of the 2019 survey and 10-m pixels, these settings yielded densities of 14.7 pulses per m^2, 18.5 returns per m^2 and 10.4 ground returns per m^2. To assess the precision of the lidar height model, the NCALM crew compared the lidar data against 965 kinematic GPS measurements processed with differential and dual wavelength geodetic techniques. The results indicate that the precision of the lidar models is within ±1.9 cm (1 s.d.) of the GPS measurements. NCALM researchers classified laser points using TerraScan software, and created a digital elevation model (DEM; a bare-earth model without vegetation and modern buildings) and a digital first surface model (including vegetation and buildings) at a horizontal spacing of 1 m for the 2017 data and 0.5 m for the 2019 data. NCALM researchers delivered the DEM and digital first surface model to the archaeologists in ESRI .flt raster format, and delivered the point cloud data in LAS format.

The examination of point clouds indicates that the high-resolution lidar used by NCALM penetrated the dense canopies of high secondary vegetation. However, where there is dense vegetation close to the ground surface (vegetation shorter than 2m (such as dense undergrowth, dense, low secondary vegetation and dense grass)), there may be mixed returns with the signals of both vegetation and the terrain. The results of our field validation suggest that, under these conditions, subtle archaeological features may be difficult to detect, but structures higher than 1.5 m can be identified in the DEM derived from the high-resolution lidar[54–56]. Most parts of our study area are covered by pasture, mature secondary vegetation or tree plantations. In these areas, low mounds and platforms—measuring 0.2 to 0.5 m in height—can usually be detected in the high-resolution lidar.

The low-resolution lidar data were collected by the INEGI (a Mexican government agency) in 2012. These data were intended for diverse uses by the Mexican government, industries, researchers of various fields and the general public. The INEGI used a Leica Geosystems ALS50-II lidar system and produced DEMs and digital first surface models at a horizontal spacing of 5 m, which are publicly available through the INEGI website (www.inegi.org.mx). The INEGI does not publish the parameters used for the acquisition of lidar data, but the laser point density appears to be generally low. We began to analyse these publicly available data in 2017. Our analysis shows that the INEGI DEMs often do not represent details of the ground topography well in forested areas. Substantial parts of our study areas, however, are deforested and used as pastures. The low-resolution INEGI lidar images show many of the large archaeological features under these conditions[57].

To examine the distribution of archaeological sites, we analysed the NCALM and INEGI lidar data using ArcGIS. We applied various visualization techniques, including hillshades, principal component analysis of multi-directional hillshades, slope gradient, sky view factor analysis, simple local relief models and red relief image map[56,58–65].

The field validation of archaeological sites is ongoing. We have visited 42 areas, which were all confirmed to be archaeological sites. In addition to Aguada Fénix and La Carmelita, five sites (Buenavista, El Macabil, El Saraguato, Rancho Zaragoza and Chrisóforo Chiñas) have been confirmed to have the MFU pattern.

Excavation

Excavations followed methods established during the investigation of Ceibal[66]. To control the proveniences of artefacts, we use a hierarchical recording system of excavation contexts, consisting of (from largest to smallest division) site code, operation, suboperation, unit, level and lot. The site codes consist of two letters: NR for the central part of Aguada Fénix; AF for peripheral areas of Aguada Fénix; LC for La Carmelita; TR for El Tiradero; and ZR for Rancho Zaragoza. An operation refers to the excavation of a mound group or a similar area; a suboperation refers to the excavation of individual structures or a small area; a unit is a horizontal division, usually of 2 × 2 m; a level is a major group of stratigraphic layers; and a lot is any natural or arbitrary division within a unit and a level. We screened all excavated soils with 1/4-inch (or smaller) mesh. We collected soil samples for floatation from important contexts (such as middens), in which we collected both floated organic materials and heavy fractions.

Middle Preclassic fills of the Aguada Fénix main plateau consisted mostly of dark clay, and floors were made of dark clay or lighter coloured earth. In operation NR3A, we identified nine Middle Preclassic floors. Thin layers of earthen fills mixed with stones were added over the Middle Preclassic construction during the Late–Terminal Preclassic (350 BC–AD 250) and the Late Classic (AD 600–810) periods. The results of operation NR7A suggest that most platforms placed along the edges of the main plateau were constructed during the Middle Preclassic period—probably before 800 BC—with earthen fills. Operations NR4A and NR6A showed that the south and west causeways were built between 950 and 800 BC with 19 to 25 successive floors, reaching total fill thicknesses of around 2.6 m.

Ceramic analysis

Because the ceramics of Aguada Fénix and La Carmelita were similar to those from Ceibal, we began our ceramic analysis by applying the ceramic typology of Ceibal[67–69]. We used Ceibal type names (such as Abelino Red, Hueche White and Crisanto Black) for ceramics that exhibited close similarities to those of Ceibal. We gave preliminary type and group names to ceramics unique to the region. They include the Tiradero group, which is characterized by thin buff to white pastes with volcanic ash temper. Some Tiradero vessels have red paint. Only a very small portion of the ceramics appears to have some affinities with materials from the Gulf Coast or Chiapas. We placed those ceramics in temporary categories. We will decide whether we will use type names from the Gulf Coast or Chiapas or whether we give new type names after we conduct thorough comparative studies with materials from other regions. We also conducted modal analysis, particularly focusing on vessel forms. Modal data also helped us to correlate the occupation of the Middle Usumacinta region with ceramic phases of Ceibal and other lowland Maya sites. We have yet to give phase names to the occupation of Aguada Fénix and La Carmelita: we will do so after we obtain more excavation data from various sites in the region.

Auger tests

We first used a hand-operated bucket auger, following the method used in the Olmec area[70]. However, it was difficult to penetrate through limestone cobbles, which are often present in the upper layers of the Aguada Fénix plateau. We then contracted a mechanical auger, which is generally used for digging wells in the region. We used a Deeprock hydraulic rotary auger DR20, which was equipped with a 4-inch point made of tungsten carbide drill tips and with metal tubes of 2-inch diameter and 5-feet length. The auger was powered by a gasoline motor,

and bored holes of 11-cm diameter. A water pump supplied water to the drill point, which extracted excavated materials. Although it is possible to use compressed air instead of water, this method was substantially more expensive. We thus decided to use the hydraulic auger.

By collecting the materials extracted with the water with a fine mesh, we could gain a general understanding of the stratigraphy as the auger advanced. The auger penetrated soft limestone blocks, but it had difficulty in penetrating hard crystallized carbonate rock or large nodules of chert. The bedrock of the area generally consists of a thin layer of soft, white marl that overlies hard carbonate rock. When the auger reached this sequence of soft and hard materials, we interpreted it as bedrock. When we encountered hard materials at depths shallower than expected, we excavated 1 × 1-m test units to verify whether we had reached bedrock. At auger test 2, we found that the auger was blocked by a large nodule of chert; at auger test 6, we confirmed that bedrock was at a depth of 1.5 m. At auger test 11 (placed on the west plateau), we reached soft, white material at depths of 7.0 and 15.0 m and hard material at 19.5 m, which made the interpretation of stratigraphy difficult.

Our stratigraphic interpretations based on the auger tests are tentative, and need to be verified with future excavations. Nonetheless, these interpretations serve the purpose of avoiding an overestimation of construction volume. Although it was sometimes difficult to determine whether soft, white layers represented the beginning of bedrock or materials included in fills, black clay layers could be reasonably interpreted as construction fills. In other words, there is the possibility that future research could reveal deeper bedrock surfaces (leading to larger estimates of construction volumes), but it is less likely that our current volume estimates become substantially smaller.

Volume calculation

Using stratigraphic data obtained from the excavations and the auger tests, we estimated the fill volume of the main plateau of Aguada Fénix. We followed the method that was used in the analysis of the plateau of Ceibal[4]. To summarize in brief, we created a three-dimensional (3D) model of the bedrock, using the Microstation CAD program. We first drew the positions of the bedrock that were found in excavations and auger tests. We then drew areas between them by assuming a smooth surface of the bedrock. For those areas, we made three versions of estimated bedrock positions: (1) the estimate that we think most likely; (2) the highest probable positions; and (3) the lowest probable positions. The 3D data of the bedrock were then imported into ArcGIS. We used the DEM derived from the NCALM high-resolution lidar as an approximation of the final form of the plateau. By subtracting the bedrock model raster files from the DEM raster, we obtained the most likely, high and low estimates of 3,790,000, 4,480,000 and 3,390,000 m^3, respectively, for the total plateau fill volume. In many areas of the plateau, we encountered fills dating to the Late–Terminal Preclassic or Classic period that measured 0.1 to 0.5 m in thickness. By using 0.3 m as an average thickness of these later constructions, we estimated the fill volume for the Late–Terminal Preclassic and Classic periods at 160,000 m^3. By subtracting this amount from the total estimated volumes, we reached the most likely, high and low estimates of 3,630,000, 4,320,000 and 3,230,000 m^3, respectively, for the Middle Preclassic fill volume (Extended Data Fig. 9b).

We determined that the effects of lidar measurement errors on these calculations are minimal, and we did not incorporate them in our volume estimates. The error range of ±1.9 cm in the NCALM lidar height model is negligible compared to the level of uncertainty in the estimates of bedrock positions. In addition, the positions of bedrock in our 3D models were plotted relative to the lidar-derived DEM, and, thus, vertical errors in lidar do not affect volume estimates in any meaningful way. Other potential factors that might affect the volume estimates include: (1) mixed returns of lidar caused by dense, low vegetation; and (2) soil erosion that happened after the abandonment of the site. There are areas of mixed returns around the east wing and the southern end of the plateau. Their total area measures 152,000 m^2. Examinations of the DEM and point clouds, as well as observations during a pedestrian survey, suggest that mixed returns may have caused the DEM to be an average of 0.1 m higher than the real ground surface in those areas. These errors may have increased a plateau volume estimate by 15,200 m^3, which is a fairly small effect. We do not have data with which to assess the quantity of soil erosion. We simply assumed that the volume loss caused by soil erosion offsets the addition by mixed returns of lidar.

The west plateau was explored with only one auger test, and its construction sequence is not clear. The auger reached possible bedrock, consisting of soft limestone or marl, at depths of 7.0 m and 15.0 m. It also hit hard rock at a depth of 19.5 m. However, this level is lower than the current surrounding ground surface, and we suspect that it is below the bedrock surface. If the bedrock surface is at 7.0 m, the volume of the west plateau would be roughly 600,000 m^3. Alternatively, the depth of 15.0 m would indicate a volume of 1,100,000 m^3.

Although calculations of volumes can contain substantial margins of error, the estimates for the main plateau are considerably larger than the volume of 2,800,000 m^3 estimated for the La Danta complex at El Mirador, the largest building complex previously known for the Maya lowlands[7] (Extended Data Fig. 9c). In addition, the estimate for the La Danta complex assumed that the underlying bedrock surface was flat. Because many large buildings in the Maya area were constructed on naturally elevated locations (as in the case of the main plateau of Aguada Fénix, and the group A plateau of Ceibal), this figure for the La Danta complex may be an overestimate. It is unlikely that the real volume of the main plateau of Aguada Fénix is smaller than that of the La Danta complex of El Mirador.

Extended Data Figure 9b lists estimates of labour investment, corresponding to different estimates of volume. Detailed methods of calculating the labour investment have been discussed in a previous publication[4]. Our study followed previous research by other scholars (including experimental work), and assumed that the plateau of Aguada Fénix is made mostly of earth[24,71–73]. For the procurement of construction materials, we used a value of 2.6 m^3 of earth dug by one person a day[71]. For the transport of materials, we used an average transport distance of 500 m and assumed that a worker carried 500 kg or 0.384 m^3 of earth a day[71]. Plateau fills contained small iron and manganese oxide nodules, which suggests that they were taken from redoximorphic soils located nearby[74]. We think that the reservoirs found west of the plateau were originally burrows that were the result of the extraction of construction material. In addition, builders possibly invested some labour in the construction of fills beyond simply dumping transported earth. However, except for the fills with coloured clays, labour investments in the construction of most fills appear to have been small. To avoid an overestimation of labour investment, we did not include labour for fill construction. Such an estimate of labour investment may have a substantial margin of error. Our purpose is to give a general idea about how many builders could have participated, and to begin to think about the social processes associated with the construction of the plateau.

Radiocarbon dating

The 69 radiocarbon samples from Aguada Fénix and La Carmelita were analysed at the University of Tokyo Radiocarbon Dating laboratory (Supplementary Table 1). Most samples were treated with the acid–alkali–acid method, but three samples with low carbon contents (TKA-21334, TKA-21339 and TKA-21344) were treated with acid only. In addition, three more samples (TKA-21330, TKA-21336 and TKA-21337) had carbon contents lower than 10%. These six samples appear to have consisted mainly of soil organic matter rather than wood charcoal, and gave dates older than other samples. Those radiocarbon dates were treated as anomalous dates.

We conducted the Bayesian analysis of radiocarbon dates using the OxCal 4.3 program and the IntCal13 calibration curve[75–78]. For studies in the Maya region, some scholars recommend mixing IntCal

(which primarily represents conditions in the northern hemisphere), with SHCal (which represents the southern hemisphere)[79,80]. However, we do not have sufficient data to understand atmospheric mixing in the region, and we decided to use IntCal13 alone, which is based on higher-quality calibration data. In addition, chronologies of many Mesoamerican sites are based on IntCal, and the use of IntCal thus facilitates chronological comparisons between different regions of Mesoamerica.

Methods of Bayesian analysis have been discussed in detail[68,81–85]: here we present a brief summary. Bayesian analysis serves to refine radiocarbon dates by incorporating stratigraphic information and other archaeological data. It also estimates the beginning and end dates for an occupation phase. Moreover, Bayesian analysis helps to identify problematic dates through the visual representation of probability distributions and statistical measures (agreement indices and outlier models). These problematic dates are excluded from subsequent Bayesian models as outliers. For a radiocarbon date with an agreement index below 60%, we need to consider the possibility that it is an outlier. Whereas agreement indices facilitate the manual rejection of outliers, outlier models statistically identify probable outliers[86]. In examining radiocarbon dates from our excavations, we made separate Bayesian models for individual operations, incorporating information on stratigraphic sequences as a prior (Supplementary Methods). Because we are in the process of building a ceramic chronology for this region, we did not incorporate ceramic sequences in the Bayesian models.

In our primary Bayesian model (model 1), we manually rejected outliers, considering contextual information and agreement indices. At Aguada Fénix and other Mesoamerican sites, problematic dates often result from the recycling of old construction materials and the stratigraphic redeposition of old construction fills. In these cases, carbon samples give radiocarbon dates older than the dates of their final depositions. Stratigraphic mixing of younger carbons through animal burrows and root growths can occur, but such cases are less frequent. Thus, when inconsistencies among stratigraphically related radiocarbon dates existed, we usually assumed that radiocarbon dates older than expected dates were outliers. In addition to model 1, we created an outlier model (model 2). The results of the two models are generally consistent, which confirms the robustness of the models. Extended Data Figure 7 presents the main results of model 1, and the complete results of model 1 are shown in Supplementary Data and Supplementary Table 1.

Six radiocarbon dates from the deposit found in operation NR3A suggest that the use of ceramics at this site started around 1250 BC (1300–1130 BC at 95.4% level). The sequence of operation NR3A also indicates that the construction of the main plateau started around 1050 BC (1130–980 BC). Bayesian model 1 gives a slightly later date for the beginning of construction at operation NR7A (1070–925 BC), but this may be because of the small number of radiocarbon dates from this excavation. Although we favour the date around 1050 BC as a conservative estimate for the beginning of plateau construction, there remains the possibility that the construction started earlier. It is not clear whether the earliest deposits found on bedrock in operations NR3A and NR7A represent middens or construction fills. These deposits contained considerable quantities of partial ceramic vessels, large sherds, shells and bones, mixed in sticky black clay. Layers of similar black clay—although with lower densities of artefacts—were found on bedrock in other excavation units across the main plateau. Although we tentatively think that the earliest deposits in operations NR3A and NR7A were placed before the initial construction of the plateau, the nature of these layers should be further investigated.

In addition, the beginning of construction in the area around the E group is not clear. Sample TKA-20670, taken from the lowest layer (under floor 23) of operation NR5A in the E-group plaza, yielded one of the earliest dates at Aguada Fénix (1385–1135 BC). For now, we tentatively assume that this context represents occupation before plateau construction or a natural soil layer. In operation NR8A (placed to the west of the E group), we did not reach bedrock. Samples TKA-21370 and TKA-21371, collected from floor 19 of this excavation, returned modelled dates of 1090–980 BC and 1095–980 BC; Bayesian model 1 gives an estimate of 1965–945 BC for the beginning of the sequence at this location. With the currently available data, we cannot determine whether TKA-21370 and TKA-21371 resulted from old wood. Thus, there is the possibility that the area around the E group was constructed earlier than the southern and northern portions of the main plateau (thus, before 1050 BC). This possibility needs to be examined with more excavations.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The results of field investigations and laboratory analyses are described more in detail in annual reports presented to the Instituto Nacional de Antropología e Historia. Those reports, as well as the 3D models for volume calculation, are available at the University of Arizona Campus Repository (https://repository.arizona.edu/handle/10150/635527).

Code availability

The OxCal code used for Bayesian analysis is provided in the Supplementary Information.

38. Chase, A. F. et al. Airborne LiDAR, archaeology, and the ancient Maya landscape at Caracol, Belize. *J. Archaeol. Sci.* **38**, 387–398 (2011).
39. Chase, A. F., Chase, D. Z., Fisher, C. T., Leisz, S. J. & Weishampel, J. F. Geospatial revolution and remote sensing LiDAR in Mesoamerican archaeology. *Proc. Natl Acad. Sci. USA* **109**, 12916–12921 (2012).
40. Chase, A. F. et al. Ancient Maya regional settlement and inter-site analysis: the 2013 west-central Belize LiDAR survey. *Adv. Archaeol. Pract.* **6**, 8671–8695 (2014).
41. Rosenswig, R. M., López-Torrijos, R., Antonelli, C. E. & Mendelsohn, R. R. Lidar mapping and surface survey of the Izapa state on the tropical piedmont of Chiapas, Mexico. *J. Archaeol. Sci.* **40**, 1493–1507 (2013).
42. Rosenswig, R. M. & López-Torrijos, R. Lidar reveals the entire kingdom of Izapa during the first millennium BC. *Antiquity* **92**, 1292–1309 (2018).
43. Hutson, S. R., Kidder, B., Lamb, C., Vallejo-Cáliz, D. & Welch, J. Small buildings and small budgets: making lidar work in northern Yucatan, Mexico. *Adv. Archaeol. Pract.* **4**, 268–283 (2016).
44. Reese-Taylor, K. et al. Boots on the ground at Yaxnohcah: ground-truthing lidar in a complex tropical landscape. *Adv. Archaeol. Pract.* **4**, 314–338 (2016).
45. Loughlin, M. L., Pool, C. A., Fernandez-Diaz, J. C. & Shrestha, R. L. Mapping the Tres Zapotes Polity: the effectiveness of lidar in tropical alluvial settings. *Adv. Archaeol. Pract.* **4**, 301–313 (2016).
46. Magnoni, A. et al. Detection thresholds of archaeological features in airborne lidar data from central Yucatán. *Adv. Archaeol. Pract.* **4**, 232–248 (2016).
47. Inomata, T. et al. Archaeological application of airborne LiDAR to examine social changes in the Ceibal region of the Maya lowlands. *PLoS ONE* **13**, e0191619 (2018).
48. Canuto, M. A. et al. Ancient lowland Maya complexity as revealed by airborne laser scanning of northern Guatemala. *Science* **361**, eaau0137 (2018).
49. Beach, T. et al. Ancient Maya wetland fields revealed under tropical forest canopy from laser scanning and multiproxy evidence. *Proc. Natl Acad. Sci. USA* **116**, 21469–21477 (2019).
50. Evans, D. H. et al. Uncovering archaeological landscapes at Angkor using lidar. *Proc. Natl Acad. Sci. USA* **110**, 12595–12600 (2013).
51. Evans, D. Airborne laser scanning as a method for exploring long-term socio-ecological dynamics in Cambodia. *J. Archaeol. Sci.* **74**, 164–175 (2016).
52. Fernandez-Diaz, J. C. et al. Capability assessment and performance metrics for the Titan multispectral mapping lidar. *Remote Sens.* **8**, 936 (2016).
53. Fernandez-Diaz, J. C., Carter, W. E., Shrestha, R. L. & Glennie, C. L. Now you see it … now you don't: understanding airborne mapping LiDAR collection and data product generation for archaeological research in Mesoamerica. *Remote Sens.* **6**, 9951–10001 (2014).
54. Hutson, S. R. Adapting LiDAR data for regional variation in the tropics: a case study from the Northern Maya lowlands. *J. Archaeol. Sci.* **4**, 252–263 (2015).
55. Prufer, K. M., Thompson, A. E. & Kennett, D. J. Evaluating airborne LiDAR for detecting settlements and modified landscapes in disturbed tropical environments at Uxbenká, Belize. *J. Archaeol. Sci.* **57**, 1–13 (2015).
56. Inomata, T. et al. Archaeological application of airborne LiDAR with object-based vegetation classification and visualization techniques at the lowland Maya site of Ceibal, Guatemala. *Remote Sens.* **9**, 563 (2017).
57. Venter, M. L., Shields, C. R. & Ordóñez, M. D. C. Mapping Matacanela: the complementary work of LiDAR and topographical survey in southern Veracruz, Mexico. *Anc. Mesoam.* **29**, 81–92 (2018).

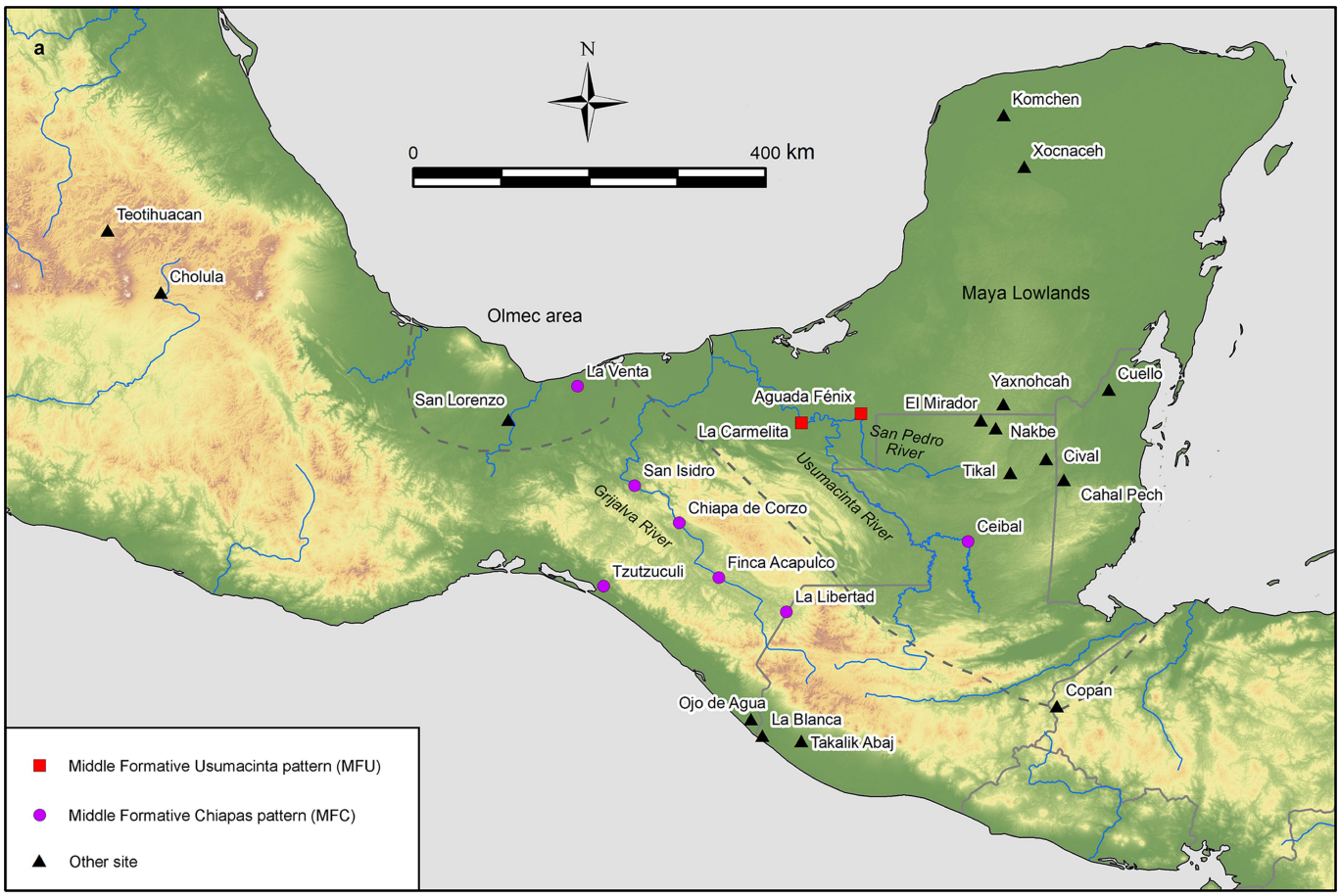
58. Bennett, R., Welham, K. & Ford, A. A comparison of visualization techniques for models created from airborne laser scanned data. *Archaeol. Prospect*. **19**, 41–48 (2012).
59. Challis, K., Forlin, P. & Kincey, M. A generic toolkit for the visualization of archaeological features on airborne LiDAR elevation data. *Archaeol. Prospect*. **18**, 279–289 (2011).
60. Devereux, B. J., Amable, G. S. & Crow, P. Visualisation of LiDAR terrain models for archaeological feature detection. *Antiquity* **82**, 470 (2008).
61. Harmon, J. M., Leone, M. P., Prince, S. D. & Snyder, M. Lidar for archaeological landscape analysis: a case study of two eighteenth-century Maryland plantation sites. *Am. Antiq*. **71**, 649–670 (2006).
62. Millard, K., Burke, C., Stiff, D. & Redden, A. Detection of a low-relief 18th-century British siege trench using LiDAR vegetation penetration capabilities at Fort Beauséjour–Fort Cumberland National Historic Site, Canada. *Geoarchaeology* **24**, 576–588 (2009).
63. Štular, B., Kokalj, Ž., Oštir, K. & Nuninger, L. Visualization of lidar-derived relief models for detection of archaeological features. *J. Archaeol. Sci*. **39**, 3354–3360 (2012).
64. Chiba, T., Kaneta, S. & Suzuki, Y. Red relief image map: new visualization method for three dimensional data. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci* **37**, 1071–1076 (2008).
65. Chiba, T. & Suzuki, Y. Visualization of airborne laser mapping data: production and development of red relief image map. *Adv. Survey Technol. (in Japanese)* **96**, 32–42 (2008).
66. Inomata, T., Triadan, D. & Aoyama, K. After 40 years: revisiting Ceibal to investigate the origins of lowland Maya civilization. *Anc. Mesoam*. **28**, 187–201 (2017).
67. Sabloff, J. A. *Excavations at Seibal, Department of Peten, Guatemala: Ceramics* (Harvard Univ., 1975).
68. Inomata, T. et al. High-precision radiocarbon dating of political collapse and dynastic origins at the Maya site of Ceibal, Guatemala. *Proc. Natl Acad. Sci. USA* **114**, 1293–1298 (2017).
69. Inomata, T. The emergence of standardized spatial plans in southern Mesoamerica: chronology and interregional interactions viewed from Ceibal, Guatemala. *Anc. Mesoam*. **28**, 329–355 (2017).
70. Cyphers, A. & Murtha, T. in *Mesoamerican Plazas: Arenas of Community and Power* (eds Tsukamoto, K. & Inomata, T.) 71-89 (Univ. Arizona Press, 2014).
71. Erasmus, C. J. Monument building: some field experiments. *Southwest. J. Anthropol*. **21**, 277–301 (1965).
72. Abrams, E. M. *How the Maya Built their World: Energetics and Ancient Architecture* (Univ. Texas Press, 1994).
73. Ortmann, A. L. & Kidder, T. R. Building mound A at Poverty Point, Louisiana: monumental public architecture, ritual practice, and implications for hunter-gatherer complexity. *Geoarchaeology* **28**, 66–86 (2013).
74. Beach, T. et al. Stability and instability on Maya lowlands tropical hillslope soils. *Geomorphology* **305**, 185–208 (2018).
75. Bronk Ramsey, C. Radiocarbon calibration and analysis of stratigraphy: the OxCal program. *Radiocarbon* **37**, 425–430 (1995).
76. Bronk Ramsey, C. Bayesian analysis of radiocarbon dates. *Radiocarbon* **51**, 337–360 (2009).
77. Bronk Ramsey, C. *OxCal 4.3*. http://c14.arch.ox.ac.uk/ (2019).
78. Reimer, P. J. et al. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal bp. *Radiocarbon* **55**, 1869–1887 (2013).
79. Kennett, D. J. et al. Correlating the ancient Maya and modern European calendars with high-precision AMS ^{14}C dating. *Sci. Rep*. **3**, 1597 (2013).
80. Hogg, A. G. et al. SHCal13 Southern Hemisphere calibration, 0–50,000 years cal BP. *Radiocarbon* **55**, 1889–1903 (2013).
81. Inomata, T., Ortiz, R., Arroyo, B. & Robinson, E. J. Chronological revisions of Preclassic Kaminaljuyú, Guatemala: implications for social processes in the southern Maya area. *Lat. Am. Antiq*. **25**, 377–408 (2014).
82. Buck, C. E., Kenworthy, J. B., Litton, C. D. & Smith, A. F. M. Combining archaeological and radiocarbon information: a Bayesian approach to calibration. *Antiquity* **65**, 808–821 (1991).
83. Buck, C. E., Cavanagh, W. G. & Litton, C. D. *Bayesian Approach to Interpreting Archaeological Data* (Wiley, 1996).
84. Bayliss, A. Rolling out revolution: using radiocarbon dating in archaeology. *Radiocarbon* **51**, 123–147 (2009).
85. Bayliss, A. Quality in Bayesian chronological models in archaeology. *World Archaeol*. **47**, 677–700 (2015).
86. Bronk Ramsey, C. Dealing with outliers and offsets in radiocarbon dating. *Radiocarbon* **51**, 1023–1045 (2009).
87. Coe, W. R. (ed). *Tikal Report No. 14: Excavations in the Great Plaza, North Terrace and North Acropolis of Tikal* (Univ. Museum, Univ. Pennsylvania, 1990).
88. Millon, R. The beginnings of Teotihuacan. *Am. Antiq*. **26**, 1–10 (1960).
89. Marquina, I. *Proyecto Cholula* (Instituto Nacional de Antropología e Historia, 1970).

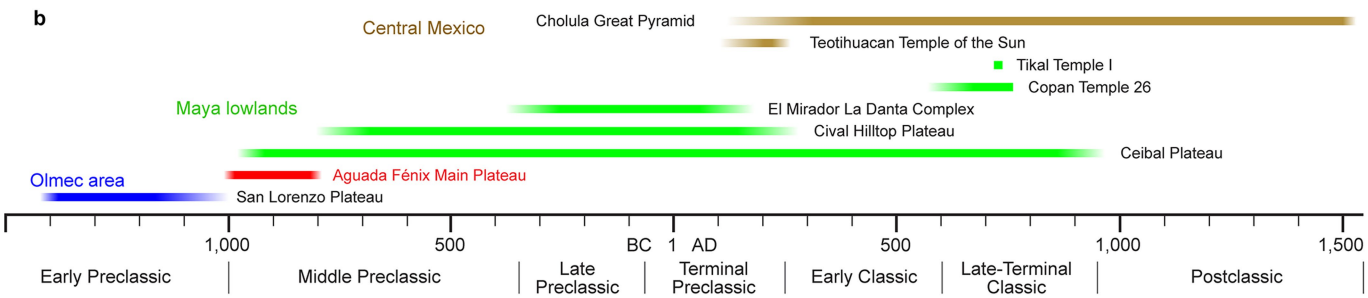
Acknowledgements The permit for our research was granted by the Instituto Nacional de Antropología e Historia. Funding was provided by the Alphawood Foundation, the National Science Foundation (BCS-1826909), the Agnese Nelms Haury Program of the University of Arizona and JSPS KAKENHI (26101003). We thank R. Liendo, K. Teranishi, F. Kupprat, V. Poston, A. Flores, F. Pinzón, M. Mollinedo, C. Alvarado, H. Zanotto, D. Ramírez, S. Mendoza and O. García for their dedicated work.

Author contributions T.I. and D.T. designed the research. T.I., D.T., M.B.M.B., V.A.V.L. and M.G.H. conducted field investigations. T.I. planned the lidar survey, and J.C.F.-D. coordinated the acquisition and processing of high-resolution lidar data. T.I. analysed lidar data and made 3D bedrock models. H.N. analysed botanical remains, and T.O. conducted radiocarbon analysis. T.I. and T.O. carried out the Bayesian analysis of radiocarbon dates. T.B. conducted soil studies, and C.C. analysed starch grains. K.A. analysed obsidian artefacts. T.I. wrote the manuscript with input from others.

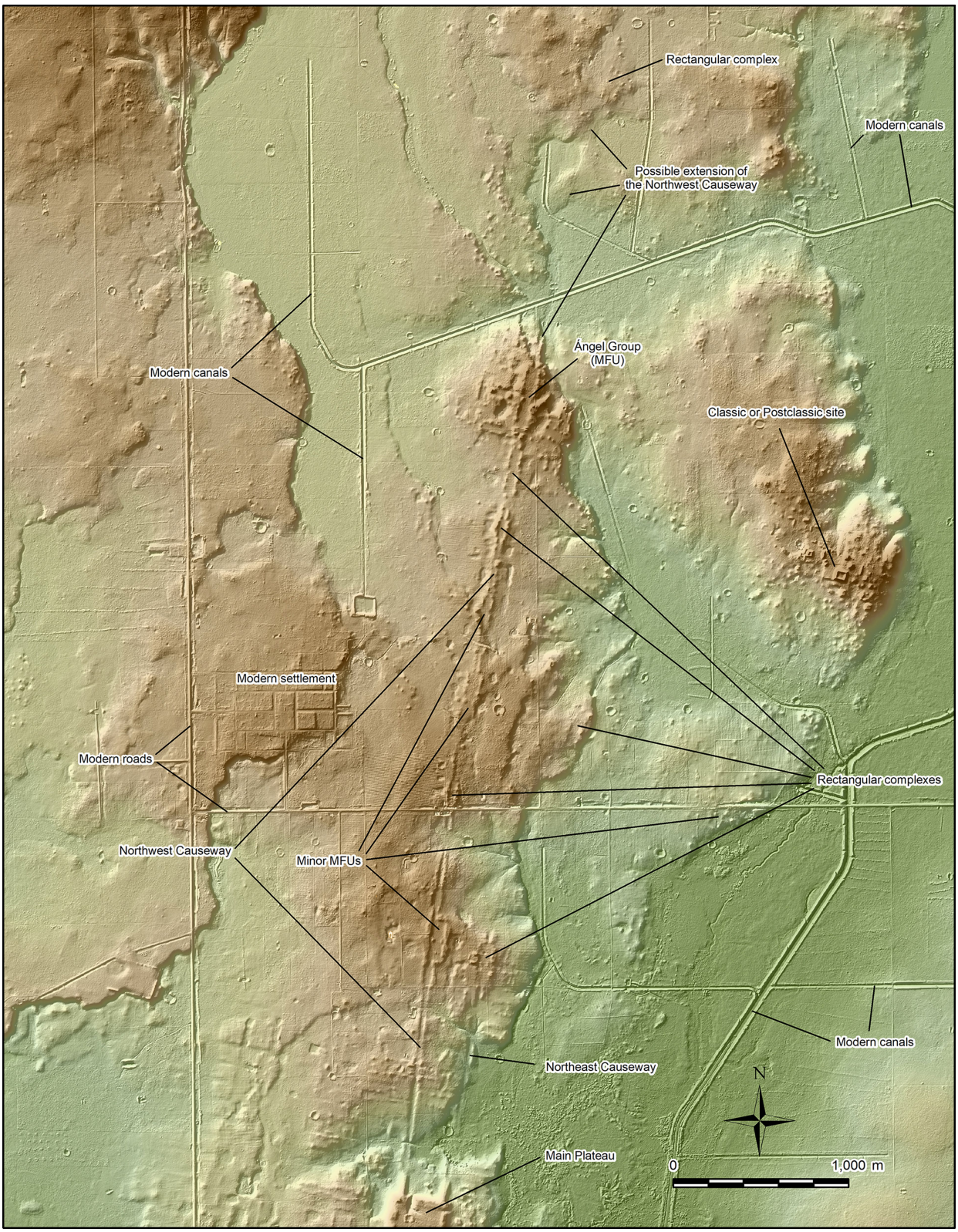
Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2343-4.
Correspondence and requests for materials should be addressed to T.I.
Reprints and permissions information is available at http://www.nature.com/reprints.

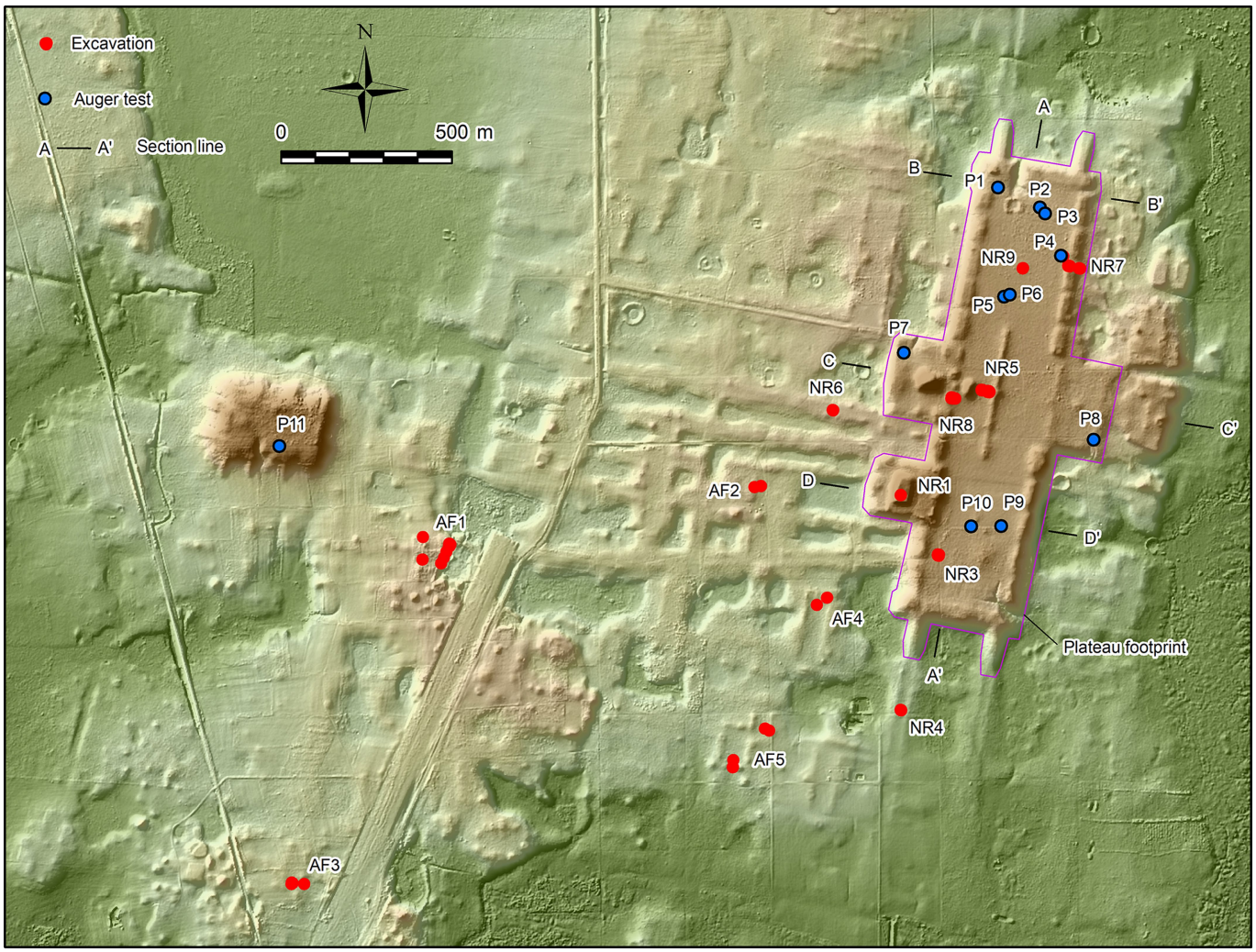




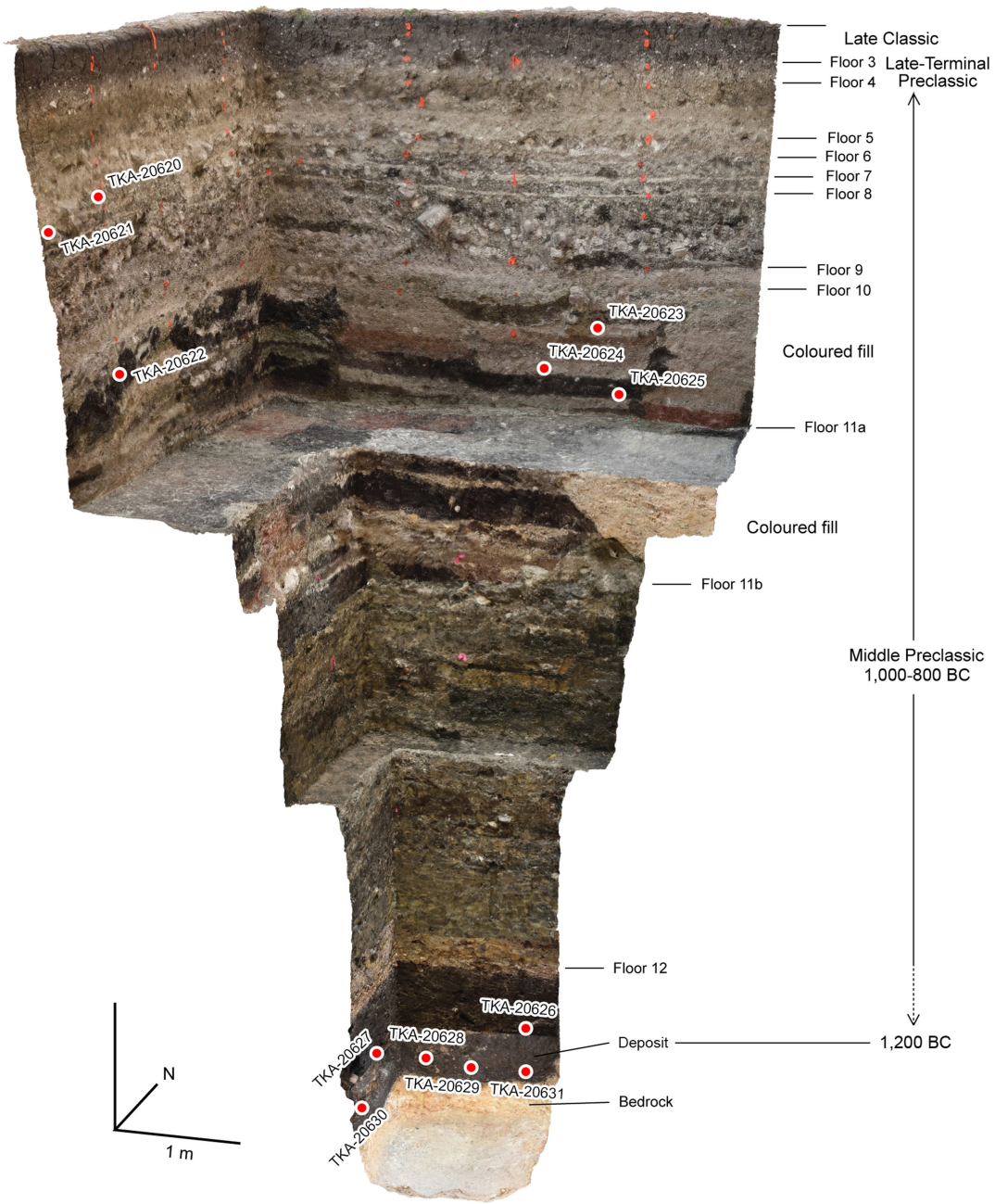
Extended Data Fig. 1 | Geographical and chronological contexts of the study. a, Map of Mesoamerica, showing the locations of the sites mentioned in the text. Map topographic data from the NASA-JPL Shuttle Radar Topographic Mission (https://www2.jpl.nasa.gov/srtm/). **b**, Chronology of Mesoamerica, indicating the construction dates of the Aguada Fénix main plateau and other major buildings listed in Extended Data Fig. 9c. Each bar shows the period in which a large portion of the building was constructed. Minor renovations and additions occurred outside of the indicated ranges.



Extended Data Fig. 2 | High-resolution lidar image of the north causeways of Aguada Fénix. The causeways are connected to the main plateau by large ramps. The northwest causeway is the longest at the site, and connects multiple MFU complexes and rectangular complexes along the way.



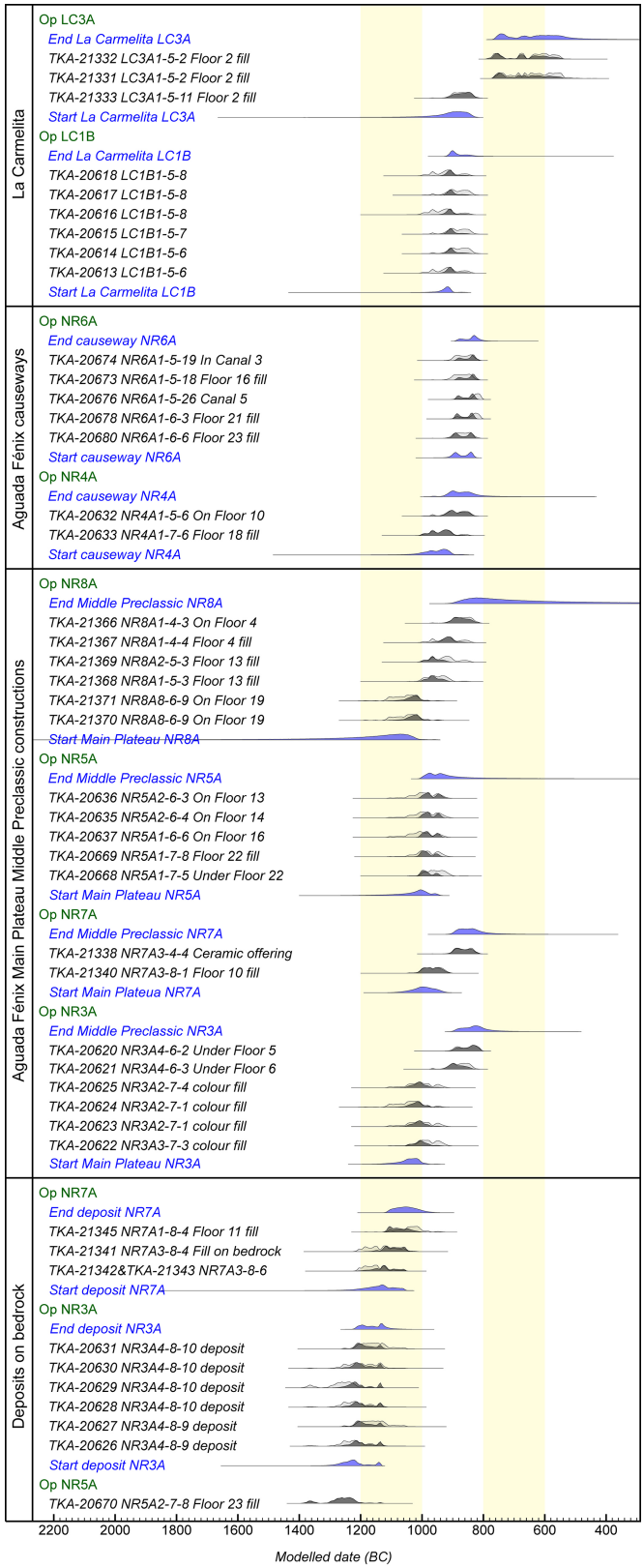
Extended Data Fig. 3 | Locations of excavations and auger tests at Aguada Fénix. The footprint of the main plateau indicated in this figure was used for the calculations of plateau fill volumes. The locations of the section drawings shown in Extended Data Fig. 9 are also indicated.



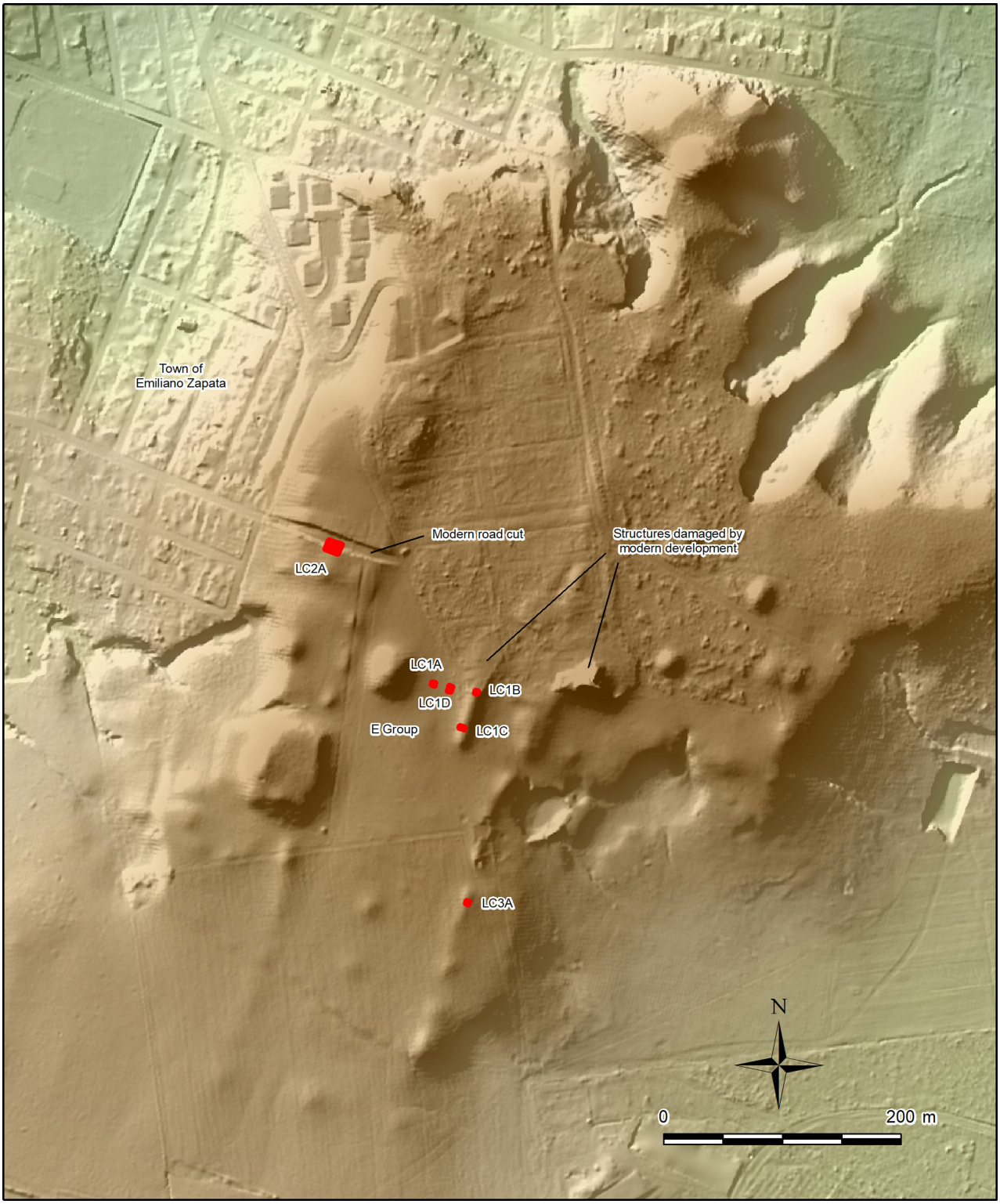
Extended Data Fig. 4 | Composite 3D photogrammetry image of operation NR3, showing the north and east profiles. The locations of radiocarbon samples are projected to the nearest profiles. The image shows that a substantial part of the plateau fills was placed during the period between 1000 and 800 BC. The fills between floors 10 and 11b consist of clays and other soils of multiple colours in checkerboard-like patterns.

Extended Data Fig. 5 | Construction fills with clays and other soils of multiple colours found in operation NR3 (a 4 × 4-m excavation, viewed from the south). a, Upper layer directly under floor 10. **b**, Middle layer. **c**, Lower layer. Blocks of soils in different colours are separated by dividers made of black clay and other soils. **d**, North profile. This sequence shows that blocks of soils in different colours were placed in multiple layers above floor 11a in one construction event. They were covered by floor 10 at the end of the sequence.

Extended Data Fig. 6 | Megalithic structure found in operation NR8. **a**, Composite 3D photogrammetry image of the structure and the excavation. **b**, Back wall viewed from the interior (from the southwest). **c**, Back wall viewed from the exterior (from the east) (2-m-wide trench). There was a deposit of broken ceramics placed at the end of the Late Classic period. **d**, Back terrace retaining wall, viewed from the east (2-m-wide trench).

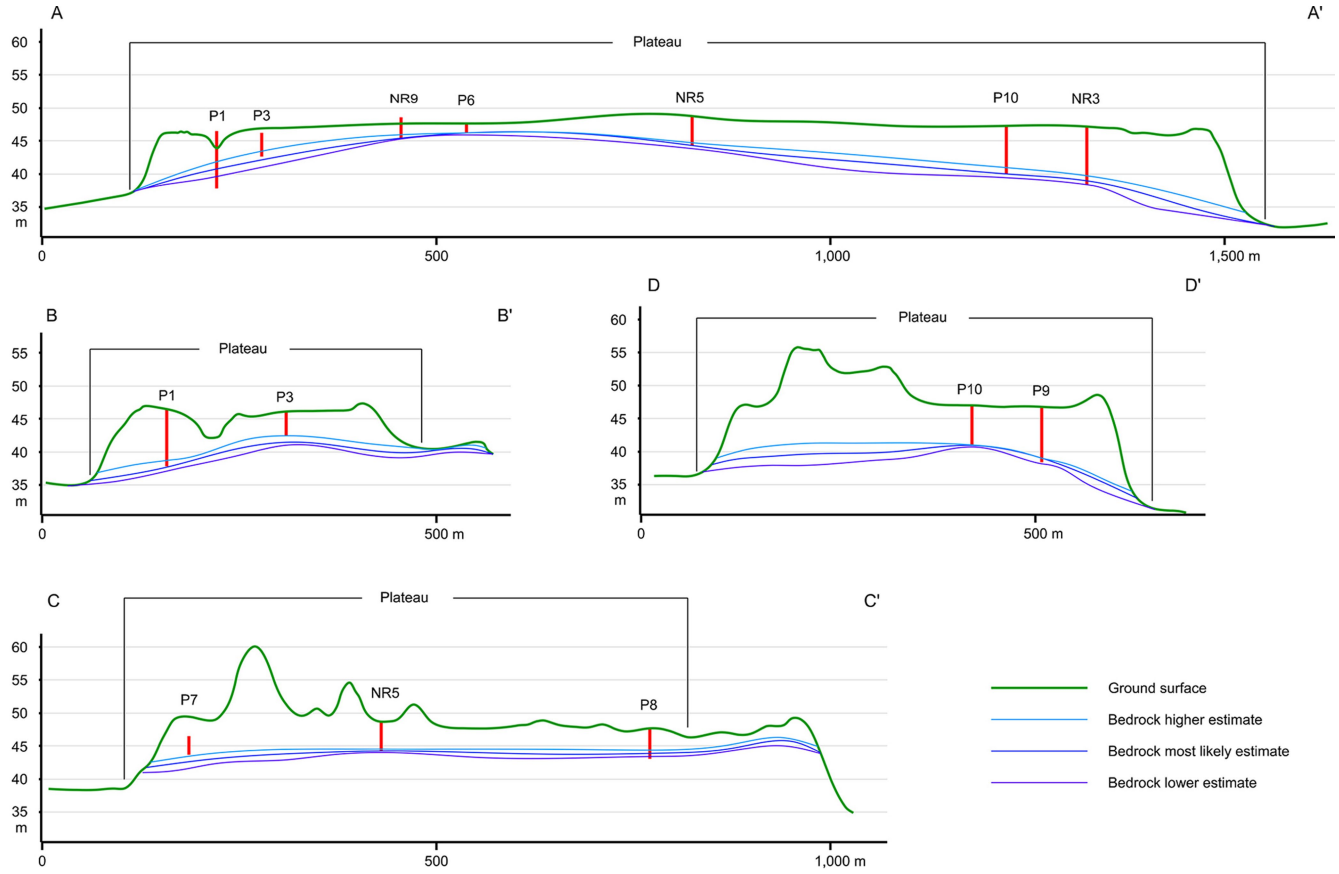


Extended Data Fig. 7 | Radiocarbon dates from Aguada Fénix and La Carmelita. Radiocarbon dates for the Middle Preclassic period and key boundary dates are shown, excluding outliers. Black areas indicate the probability distributions of modelled dates obtained with model 1, and grey areas show those of unmodelled calibrated dates. Dates in blue represent boundary dates. The entire OxCal results of model 1 are provided in Supplementary Table 1 and Supplementary Data.



Extended Data Fig. 8 | Locations of excavations at La Carmelita. The northern part of the site, including the northern portion of the eastern platform of the E group, was damaged by a modern development project. The construction was halted by the Mexican government after initial destruction.

a



b

Aguada Fénix structure	Estimate	Volume (m^3)	Labor investment (person-day)
Main Plateau all periods	High	4,480,000	13,400,000
	Most likely	3,790,000	11,300,000
	Low	3,390,000	10,100,000
Main Plateau Middle Preclassic	High	4,320,000	12,900,000
	Most likely	3,630,000	10,800,000
	Low	3,230,000	9,700,000
West Plateau Middle Preclassic	High	1,150,000	3,400,000
	Low	610,000	1,800,000
Combined Middle Preclassic	High	5,470,000	16,300,000
	Low	3,840,000	11,500,000

c

Site	Structure	Volume (m^3)	Period	Reference
Aguada Fénix	Main Plateau	3,200,000-4,300,000	Middle Preclassic (1,000-800 BC)	
	West Plateau	600,000-1,200,000	Middle Preclassic	
	Combined	3,800,000-5,500,000	Middle Preclassic	
Ceibal	Group A Plateau	710,000	Preclassic to Classic	4
Cival	Hilltop plateau	1,900,000	Preclassic	5
El Mirador	La Danta Complex	2,800,000	Mostly Late to Terminal Preclassic	7
	El Tigre Complex	430,000	Mostly Late to Terminal Preclassic	7
	Combined	3,200,000	Mostly Late to Terminal Preclassic	7
Tikal	Temple I	18,300	Late Classic	24,87
Copan	Temple 26	31,900	Mostly Late Classic	24
San Lorenzo	Plateau	6,000,000-8,000,000	Early Preclassic	13
Teotihuacan	Pyramid of the Sun	1,600,000	Terminal Preclassic	88
	Pyramid of the Moon	320,000	Terminal Preclassic to Early Classic	88
	Combined	1,900,000	Terminal Preclassic to Early Classic	88
Cholula	Great Pyramid	4,500,000	Preclassic to Postclassic	89

Extended Data Fig. 9 | Calculation of the volume of the main plateau at Aguada Fénix. a, Section drawings of the plateau, showing the current ground surface and the estimated positions of bedrock. Vertical dimensions are exaggerated. The locations of the section lines are shown in Extended Data Fig. 3. Red lines indicate the depths of bedrock reached by excavations and auger tests. When excavations and auger tests are not on the section lines, their elevations may not correspond exactly with the positions of the current ground surface and bedrock shown here. **b**, Estimated construction volumes of the main plateau and the west plateau of Aguada Fénix, and estimates of labour investment. **c**, Comparison of the Aguada Fénix plateaus with other major buildings[4,5,7,13,24,87–89] in Mesoamerica. The construction volume of the main plateau of Aguada Fénix is larger than that of the La Danta complex (the largest construction in the Maya lowlands previously known) and that of the Pyramid of the Sun of Teotihuacan, the largest city in Preclassic-to-Classic Mesoamerica. The Great Pyramid of Cholula is larger, but it was expanded over more than 1,000 years.

Extended Data Fig. 10 | Early Middle Preclassic caches found at Aguada Fénix. a, **b**, Cache NR3 (found in operation NR5B), which was placed on the east–west axis of the E-group plaza. It contained six axes and a perforator (all made of greenstone), as well as three small pieces of greenstone. The pointed end of the perforator is broken. The contents and location of this cache closely resemble those found at San Isidro, Chiapa de Corzo, Ceibal and Cival. Similar caches of greenstone axes were also found at La Venta, although not in the E-group plaza. These deposits, along with the similarities in site layout, show that these Middle Preclassic centres shared spatial and ritual concepts. **c**–**e**, Cache AF1, found in operation AF1D. It contained a limestone sculpture—possibly representing a white-lipped peccary—that we named 'Choco'. The naturalistic image of an animal contrasts with Olmec art, which depicts supernatural beings and high-status individuals.

nature research

Corresponding author(s): Takeshi Inomata

Last updated by author(s): Apr 6, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Lidar data acquisition and processing were done with Optech LMS 4.4.0, Terrasolid TerraScan 019.003 and Golden Software Surfer 12.
Data analysis	Lidar-derived DEMs were analyzed with ESRI ArcGIS 10.7.1. The production of a 3D model of bedrock and fill volume calculation were done with ArcGIS and Bentley Microstation 08.11.09.829. The composite phogrammetry image of excavation (Extended Data Figure 4) was made with Agisoft PhotoScan 1.4.4. The Bayesian analysis of radiocarbon dates was done with Oxcal 4.3. The Oxcal codes for this analysis are included as Supplementary Information.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The results of field investigations and lab analyses are described more in detail in the annual reports presented to the Instituto Nacional de Antropología e Historia. Those reports, as well as the 3D models for volume calculation, are available at the University of Arizona Campus Repository (https://repository.arizona.edu/arizona/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This is an archaeological study of past society, including excavations, surveys, lidar, artifact analysis and radiocarbon dating. It involves quantitative data on structures sizes and radiocarbon dates, as well as the qualitative study of social processes.
Research sample	The research sample consists of archaeological data obtained from lidar, ground surveys, excavations, artifact analysis and radiocarbon dates. We used existing low-resolution lidar data, which were made publicly avaible by the INEGI (www.inegi.org.mx) and covers the entire study area. We selected the areas for high-resolution lidar, where important sites were found in the INEGI lidar. At Aguada Fénix, we selected 5 excavation areas on the plateau, 2 areas on causeways, and 5 areas in the periphery to examine construction history across the site. At La Carmelita, we selected 5 excavation areas to examine the construction history of this smaller site. We chose sixty nine radiocarbon dates to date the entire the entire occupation history of Aguada Fénix and La Carmelita.
Sampling strategy	No sample size calculation was performed. The high-resolution lidar covers the entire sites of Aguada Fénix and La Carmelita. The locations of excavations and auger tests were selected to cover different parts of the sites. Thus, the excavation samples are representative of the construction volume and occupation history. The excavations at La Carmelita provide representative data to reconstruction the construction history of its ceremonial core. Sixty nine radiocarbon dates cover the entire occupation sequences and are representative.
Data collection	The high-resolution lidar data were collected with Optech Titan lidar. Excavation data were recorded on paper forms in the field and then input in computer files. Photographs of excavations were taken with Nikon D750 and D7000 digital cameras. In addition to the authors of this paper, other archaeologists, archaeology students, and local community members participated in excavations.
Timing	Archaeological fieldwork was conducted July-August 2017, February-April 2018, February-April 2019 and February-March 2020. High-resolution lidar data were collected in May 2017 and June 2019.
Data exclusions	No excavation data were excluded. In the Bayesian analysis of radiocarbon dates, we followed the pre-established and commonly accepted criteria for exclusion (agreement indices lower than about 60 and the results of outlier models). Fourteen dates were excluded as outliers from the models.
Non-participation	The study does not involve participants.
Randomization	Locations for excavations and lidar surveys were not randomized. As the lidar covers the entire sites and excavations targeted different parts of the sites, they provided necessary data for the reconstruction of construction sequences and fill volumes.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Lineage dynamics of the endosymbiotic cell type in the soft coral *Xenia*

https://doi.org/10.1038/s41586-020-2385-7

Received: 26 June 2019

Accepted: 28 April 2020

Published online: 17 June 2020

Open access

Check for updates

Minjie Hu[1]✉, Xiaobin Zheng[1], Chen-Ming Fan[1]✉ & Yixian Zheng[1]✉

Many corals harbour symbiotic dinoflagellate algae. The algae live inside coral cells in a specialized membrane compartment known as the symbiosome, which shares the photosynthetically fixed carbon with coral host cells while host cells provide inorganic carbon to the algae for photosynthesis[1]. This endosymbiosis—which is critical for the maintenance of coral reef ecosystems—is increasingly threatened by environmental stressors that lead to coral bleaching (that is, the disruption of endosymbiosis), which in turn leads to coral death and the degradation of marine ecosystems[2]. The molecular pathways that orchestrate the recognition, uptake and maintenance of algae in coral cells remain poorly understood. Here we report the chromosome-level genome assembly of a *Xenia* species of fast-growing soft coral[3], and use this species as a model to investigate coral–alga endosymbiosis. Single-cell RNA sequencing identified 16 cell clusters, including gastrodermal cells and cnidocytes, in *Xenia* sp. We identified the endosymbiotic cell type, which expresses a distinct set of genes that are implicated in the recognition, phagocytosis and/or endocytosis, and maintenance of algae, as well as in the immune modulation of host coral cells. By coupling *Xenia* sp. regeneration and single-cell RNA sequencing, we observed a dynamic lineage progression of the endosymbiotic cells. The conserved genes associated with endosymbiosis that are reported here may help to reveal common principles by which different corals take up or lose their endosymbionts.

Many corals take up dinoflagellate algae of the Symbiodiniaceae family into their gastrodermis through feeding. Some cells in the gastrodermis, which lines the digestive tract, may have the ability to recognize particular types of algae. Through phagocytosis and by modulating host immune responses, the matching algal type is enclosed by endomembranes to form symbiosomes inside coral cells[1]. The symbiosome membrane is believed to contain transporters that mediate nutrient exchange between the algae and host cells[4]. Comparative transcriptome analyses on whole organisms using different cnidarian species before and after algae colonization or bleaching have identified genes, the up- or downregulation of which could contribute to endosymbiosis[5–7]. Comparative genomic and transcriptomic information in endosymbiotic and non-symbiotic cnidarian species has also been used to search for genes that may have evolved to mediate the recognition or endocytosis of Symbiodiniaceae[6–9]. However, these approaches do not differentiate whether the altered genes are expressed in the host endosymbiotic cells or other cell types without additional criteria. Protein inhibition or activation has also been used to suggest that host proteins containing C-type lectin domains, scavenger receptor domains or thrombospondin type 1 repeats are involved in uptake of algae and immunosuppression[10–12]. The broad expression and function of these proteins, coupled with potential off-target effects of inhibitors, greatly limit data interpretation. Therefore, a systematic description of genes and pathways that are selectively expressed in the host endosymbiotic cells is much needed to begin to understand the potential regulatory mechanisms that underlie the entry, establishment and—possibly—the expulsion of Symbiodiniaceae.

Genome and single-cell transcriptome

We chose to study a *Xenia* sp. of pulsing soft coral (Fig. 1a, b, Extended Data Fig. 1, Supplementary Video 1) that grows rapidly in a laboratory aquarium. Using Illumina short-read and Nanopore long-read sequencing (Extended Data Table 1), we assembled the *Xenia* genome into 556 high-quality contigs. Applying chromosome conformation capture (Hi-C)[13,14], we further assembled these contigs into 168 scaffolds; the longest 15 of these scaffolds contain 92.5% of the assembled genome of 222,699,500 bp, consistent with the GenomeScope estimation (Extended Data Fig. 2). To our knowledge, the *Xenia* genome has by far the longest scaffold length, and thus the most contiguous assembly, of the published cnidarian genomes (Fig. 1c). Annotation using several bulk RNA-sequencing (RNA-seq) datasets showed that *Xenia* sp. has 29,015 genes, similar to other cnidarians (Extended Data Tables 2, 3). Consistent with previous phylogenetic analyses[15], the octocorallians, *Xenia* sp., *Dendronephthya gigantea* and *Renilla reniformis* are grouped as a clade that is sister to the hexacorallian clade (which contains sea anemones and scleractinian corals), as they are all anthozoans (Fig. 1d).

We next performed single-cell RNA-seq (scRNA-seq)[16] of whole polyps, stalks or tentacles using version 2 and version 3 chemistry of the 10x Genomics platform (Supplementary Table 1, Methods). Using *t*-distributed stochastic neighbour embedding (*t*-SNE)[17], we grouped the high-quality single-cell transcriptomes, covering 23,939 genes, into 16 cell clusters with distinct gene-expression patterns (Fig. 2a, b, Extended Data Fig. 3a, Supplementary Table 2). For validation, we

[1]Department of Embryology, Carnegie Institution for Science, Baltimore, MD, USA. ✉e-mail: mhu2@carnegiescience.edu; fan@carnegiescience.edu; zheng@carnegiescience.edu

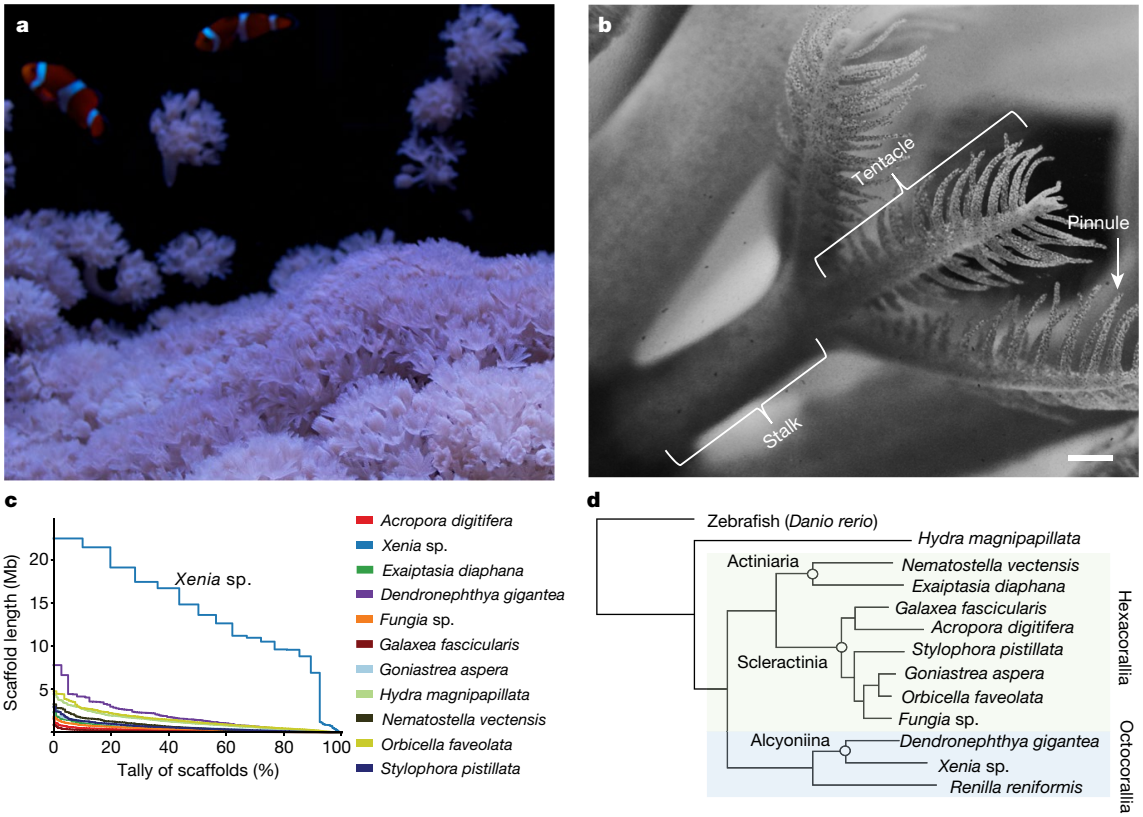


Fig. 1 | High-quality genome assembly for *Xenia* sp. a, *Xenia* sp. grown in the laboratory aquarium. **b**, An enlarged view of a *Xenia* sp. polyp with its main substructures indicated. Scale bar, 1 mm. **c**, Comparisons of the assembled scaffold lengths (*y* axis) and tallies (*x* axis) of 11 sequenced cnidarians, including *Xenia* sp. **d**, Evolutionary comparisons of *Xenia* sp. with other cnidarians, as indicated. Zebrafish and *Hydra* were used as outgroups. The phylogenetic branch points were assigned with 100% confidence.

looked for two previously characterized cnidarian cells: the cnidocytes, which are used for prey capture and/or defence, and gastrodermal cells. The cells of cluster 11 express minicollagen and nematogalectin genes, which are markers of cnidocytes[18–20] (Fig. 2c). Further analysis revealed that cluster 11 contained two subclusters (Fig. 2d, Extended Data Fig. 3b). Minicollagen genes are expressed in both subclusters, whereas nematogalectin genes are preferentially expressed in one (Fig. 2e, Extended Data Fig. 3c). RNA in situ hybridization (ISH) confirmed the expression of a nemetogalectin gene to be more spatially restricted than that of *Minicollagen 1* in *Xenia* pinnules (Fig. 2f, g, Extended Data Fig. 3d, e). Clusters 2, 12 and 16 express genes that encode collagens and proteases (Fig. 2h) that are known to be enriched in gastrodermis of *Nematostella*[18]. RNA ISH for *Collagen 6*, *Astacin-like metalloendopeptidase 2* (both expressed by clusters 2 and 12) and the uncharacterized *Xe_003623* gene (expressed by clusters 2, 12 and 16) confirmed the high expression of these genes in the gastrodermis (Fig. 2h–j, Extended Data Fig. 3f–i). Thus, the clustering analyses and ISH identified cnidocytes and cells in the gastrodermis in *Xenia*.

Endosymbiotic cell type in *Xenia* sp.

To identify the cells that perform endosymbiosis, we took advantage of the autofluorescence of the member of the Symbiodiniaceae (*Durusdinium*) in our *Xenia* sp. (Methods). Using fluorescence-activated cell sorting (FACS), we separated alga-containing and alga-free *Xenia* cells (Fig. 3a, b) and performed bulk RNA-seq (Supplementary Table 3). By comparing these bulk transcriptomes with genes expressed in each cluster, we found that cells of cluster 16 exhibited the highest overall similarity to the alga-containing cells and most of the marker genes for cluster 16 (Supplementary Table 4) have a higher level of expression in alga-containing *Xenia* cells than that in alga-free *Xenia* cells (Fig. 3c, d, Supplementary Table 5). RNAscope ISH for two of the cluster-16 marker genes—one of which encodes a protein with lectin and kazal protease inhibitor domains (abbreviated LePin, encoded by a gene that we name *LePin*), and the other of which encodes Granulin 1—showed that these genes were expressed in alga-containing gastrodermal cells (Fig. 3e, f, Extended Data Fig. 4a, b). Additionally, on average 95% and 98% of alga-containing *Xenia* cells were positive for expression of *LePin* and *Granulin 1*, respectively (Extended Data Fig. 4c). On the basis of microscopy of cryopreserved tissue sections or FACS analyses, we estimated that on average 2–6% of *Xenia* cells contained algae and that tentacles have a higher percentage of alga-containing cells than do stalks (Extended Data Fig. 4d, Methods). This is consistent with the cluster-16 endosymbiotic cells being identified by scRNA-seq as a small fraction (382 cells, 1.4% of the total). Of the three gastrodermal cell clusters, cluster-16 cells therefore have a high likelihood of being a major cell type involved in endosymbiosis.

Among the top 89 marker genes enriched in the cluster-16 endosymbiotic cells, 67 encode proteins with domains of known or predicted functions, including receptors, extracellular matrix proteins, immune response proteins, phagocytosis and/or endocytosis proteins, or nutrient transporters (Extended Data Fig. 4e, Supplementary Table 4). Three proteins—encoded by *CD36*, *DMTB1* and *CUZD1*—contain CD36 or scavenger receptor domains that are known to recognize a wide range of microbial surface ligands and mediate their phagocytosis, and that also modulate the innate immune response of the host[11,21,22] (Fig. 3g, Extended Data Fig. 5a). CUZD1 is the least understood, and is similar to DMBT1 in domain organization. DMBT1 functions in pattern recognition of microorganisms. In mammals, it is expressed on the surface of the gastrointestinal tract, where it recognizes polysulfated and polyphosphorylated ligands on microorganisms, represses the

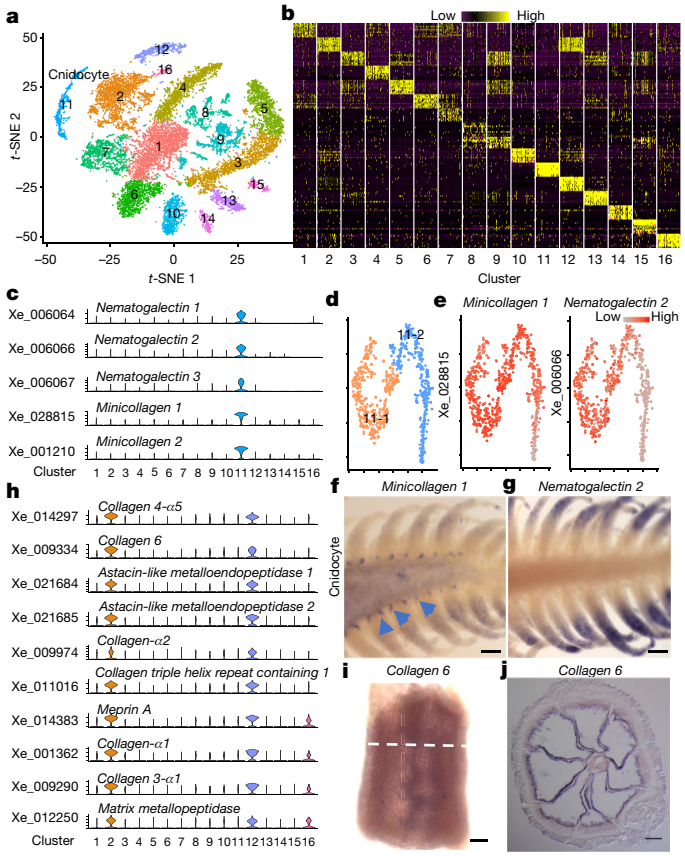


Fig. 2 | scRNA-seq transcriptomes suggest that there are 16 cell types in *Xenia* sp. a, Transcriptomes of 19,134 individual *Xenia* sp. cells obtained by scRNA-seq were grouped into 16 clusters (colour-coded) and presented in *t*-SNE space. Each coloured dot represents one cell. **b**, Gene-expression heat map (scale at the top) for the top 10 gene markers that define each cluster. Each column represents one cell cluster, and each row represents one gene. Forty cells were randomly selected from each of the 16 cell clusters for plotting. **c**, Expression profiles of the indicated cluster-11 *Xenia* (Xe) marker genes out of all cell clusters. **d**, Cluster-11 cells are subdivided into two populations (11-1, 423 cells; 11-2, 374 cells, colour-coded) and displayed in a *t*-SNE space. Each coloured dot represents a cell. **e**, Expression levels (scale to the top right) of two cluster-11 markers, *Minicollagen 1* and *Nematogalectin 2*, are shown in a *t*-SNE plot. $n = 797$ cells. **f**, **g**, Whole-mount RNA ISH of *Minicollagen 1* (**f**) and *Nematogalectin 2* (**g**), showing their expression in tentacles. Arrows indicate the expression of *Minicollagen 1* at the base of pinnules. **h**, Expression profiles of marker genes enriched in clusters 2, 12 and 16 out of all 16 clusters. **i**, **j**, RNA ISH of *Collagen 6*. Whole-mount view of the stalk in **i** and cross-section image in **j**. The white dashed line in **i** indicates the cross-section level in **j**. More than 12 polyps from 4 independent experiments were used for each probe. Scale bars, 100 µm (**f**, **g**, **j**), 150 µm (**i**). Cell numbers for clusters 1–16 are 2,794; 2,704; 2,073; 1,679; 1,511; 1,374; 1,248; 1,069; 986; 923; 797; 649; 575; 321; 246; and 185, respectively (**a**, **c**, **h**).

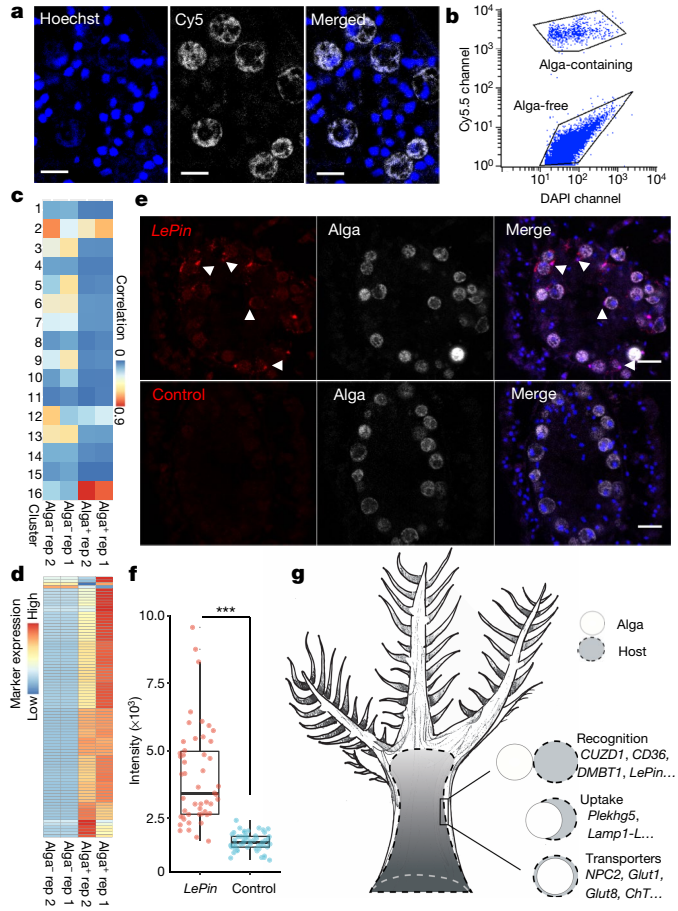


Fig. 3 | Identification of genes specifically expressed in *Xenia* sp. endosymbiotic cells. a, The endosymbiotic algae in *Xenia* display autofluorescence in the Cy5.5 far-red channel. A cross-section of *Xenia*, with the *Xenia* and algal nuclei stained by Hoechst (blue) and alga autofluorescence (white). **b**, A FACS profile of dissociated live *Xenia* cells using Cy5.5 and DAPI channels. Five biological replicates (**a**, **b**). **c**, Pearson correlation of gene expression between the scRNA-seq data of 16 cell clusters and the bulk RNA-seq data of 2 biological replicates of FACS-isolated alga-containing ($alga^+$ rep 1 and $alga^+$ rep 2) and alga-free ($alga^-$ rep 1 and $alga^-$ rep 2) cells. **d**, Heat map showing the expression levels of the 89 marker genes for cluster-16 cells in alga-containing and alga-free cells, isolated by FACS. **e**, Ultra-sensitive fluorescence RNA ISH by RNAscope probing for *LePin* (red) (top) and control (bottom). White arrows indicate the *LePin* signal. Hoechst staining of all nuclei is shown in blue. Scale bars, 20 µm. **f**, Quantification of *LePin* signals. The fluorescence signal surrounding each alga is quantified in random sections and plotted (a dot = a section) for *LePin* and controls. $^{***}P = 2.84 \times 10^{-16}$, two-sided *t*-test. Lines in the box denote the median; the upper and lower edges of the box represent the upper and lower quartiles, respectively. Nine polyps from three independent experiments were used for each probe (**e**, **f**). **g**, Illustration of steps through which *Xenia* endosymbiotic cells may recognize and take up algae to establish endosymbiosis, with some candidate genes shown at each step.

inflammatory response and regulates the differentiation of gastrointestinal cells[23]. *LePin* and *Granulin 1*, which we used for ISH, have homologues in *Exaiptasia*, as well as stony and soft corals. Because LePin has an N-terminal signal peptide followed by multiple domains (including H- and C-type lectins and a Kazal-type serine protease inhibitor) (Extended Data Fig. 5b), it may confer selectivity for the Symbiodiniaceae. On the basis of previous studies of granulins in mammals[24], *Granulin 1* may modulate the immune response in *Xenia* endosymbiotic cells.

Phagocytosis of the Symbiodiniaceae by gastrodermal cells (which are of a similar size to these algal cells) requires substantial expansion of the host cell, but the genes that regulate this size expansion are unknown. Among the endosymbitoic marker genes that we found, *Plekhg5* encodes a highly conserved RhoGEF (Fig. 3g, Extended Data Fig. 5c). In *Xenopus*, Plekhg5 localizes to the apical membrane of epithelial cells and recruits actomyosin to induce cell elongation and apical constriction[25]. Thus, *Plekhg5* is a prime candidate for regulating the extension of the apical membrane to engulf algae of the Symbiodiniaceae during the early stages of phagocytosis in *Xenia*. Upon phagocytosis, algae of the Symbiodiniaceae are enclosed by the host membrane to form symbiosomes[26]. Although the symbiosome is acidified similarly

to lysosomes[27], the genes that are involved in the formation of the symbiosome remain unclear. *Xenia* sp. has two genes that encode lysosome-associated membrane glycoproteins, which are more similar to the previously characterized LAMP1 than to LAMP2[28]. In *Xenia*, *Lamp1-L* encodes a larger protein and is an endosymbiotic marker gene, whereas *Lamp1-S* encodes a smaller protein and is expressed across all cell clusters (Extended Data Fig. 5d, e). Because lysosome-associated membrane glycoproteins are known to regulate phagocytosis, endocytosis, lipid transport and autophagy[28], Lamp1-L may regulate symbiosome formation and/or function (Fig. 3g). Several endosymbiotic marker genes encode enzymes that may promote the establishment of endosymbiosis or facilitate nutrient exchanges between alga and the host cell. For example, there are 17 genes that potentially participate in nutrient exchanges, as they encode transporters for sugar, amino acids, ammonium, water, cholesterol and choline (Fig. 3g, Extended Data Fig. 4e, Supplementary Table 4).

Lineage dynamics of endosymbiotic cells

To better understand the temporal dynamics of cluster-16 cells, we developed a *Xenia* regeneration model. We surgically cut away all tentacles from *Xenia* polyps and found that the stalks regenerated all tentacles in several days when cultured in the seawater from our aquarium that houses stock corals (Fig. 4a). Individual tentacles also regenerated into full polyps, but required a longer time (data not shown). BrdU labelling showed that some proliferated (BrdU$^+$) gastrodermal cells began to take up algae that were present either in the gastrodermis or in the seawater at day 4 of regeneration (Extended Data Fig. 6a, b). We performed scRNA-seq of the regenerating stalks and pooled the data with the scRNA-seq of non-regenerating samples (Methods).

We used Monocle 2 to perform pseudotemporal ordering of all of the endosymbiotic *Xenia* cells[29] (Fig. 4b); Monocle 2 uses reversed graph embedding to construct a principal curve that passes through the middle of the cells in the *t*-SNE space. Because this trajectory analysis does not provide a direction of cell-state progression, we used velocyto[30] to determine the directionality of lineage progression of all cells and focused on the endosymbiotic cells in the regenerating sample. Velocyto calculates RNA velocity by comparing the number of unspliced and spliced reads, which measures the expected change in gene expression in the near future—thereby providing the directionality of cell-state change. This enabled the identification of early and late stages of endosymbiotic cells (green and red, respectively, in Extended Data Fig. 6c). The cell trajectory showed that the early and late-stage cells are mapped to relatively early and late pseudotime, respectively (Extended Data Fig. 6d). Thus, the pseudotime represents actual lineage progression. Modelling of gene expression revealed substantial changes along pseudotime. Further hierarchical clustering showed distinct gene-expression patterns, which helped to define five putative endosymbiotic cell states (Fig. 4c, Supplementary Table 6).

To further explore the cell dynamics in these five states, we compared single-cell transcriptomes to transcriptomes from the bulk RNA-seq of alga-containing or alga-free cells isolated by FACS, and plotted the expression correlation along pseudotime. State-3 cells showed the strongest correlation with the alga-containing cells, followed by state 2 and then state 1; state-4 and state-5 cells showed the least correlation (Fig. 4d). This suggests that state 3 represents mature, alga-containing cells. State-1 and state-5 cells showed correlations with alga-free cells (Fig. 4d). Given that these five states are present in our identified endosymbiotic cell type with a linear pseudotime progression, we hypothesize that state-1 cells are pre-endosymbiotic progenitors that can transition through state 2 to become state-3 mature alga-containing cells, and that state-3 cells could further transit through state 4 into state-5 post-endosymbiotic cells (Fig. 4e). In support of this, we found that the regenerating samples have higher percentages

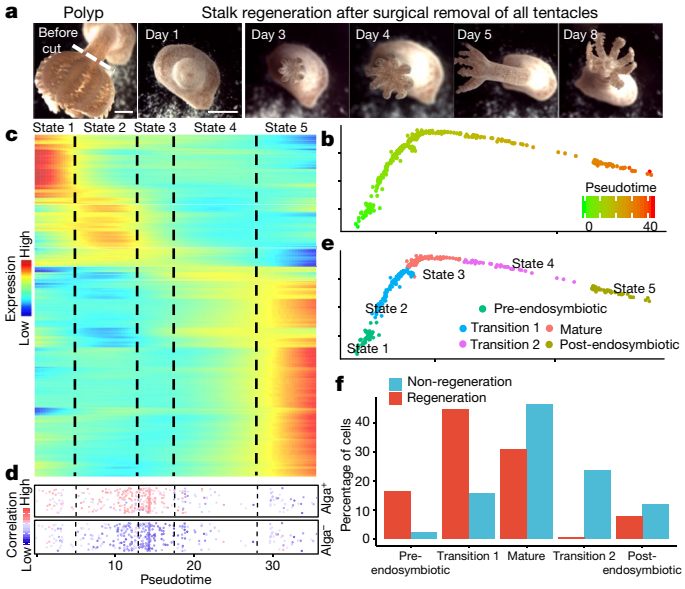


Fig. 4 | Dynamic lineage progression of endosymbiotic cells. a, An example of a *Xenia* sp. polyp (shown in the panel on the far left) is surgically cut at the white dashed line to remove all the tentacles. A surgically cut stalk is shown to regenerate in successive days as indicated. Five biological replicates. Scale bars, 1 mm. **b**, Pseudotime trajectory of all endosymbiotic cells (a dot represents a cell) identified in regeneration and non-regeneration scRNA-seq datasets. The pseudotime indicator is shown at the bottom right. **c**, Heat map for gene-expression levels along pseudotime. **d**, Correlation of the scRNA-seq transcriptome with the bulk RNA-seq transcriptome of alga-containing or alga-free *Xenia* cells, isolated by FACS. The endosymbiotic cells used to model gene expression along the pseudotime line are aligned with the heat map. Each cell represented by a dot is coloured according to the Pearson correlation of its transcriptome with the indicated bulk RNA-seq transcriptome. Five cell states (states 1–5, separated by dashed lines) are defined by differential gene expression together with the Pearson correlation. **e**, Five states defined by **c** and **d**. **f**, The percentage of cells in each state in regeneration and non-regeneration samples.

of state-1 (pre-endosymbiotic) and state-2 (transition 1) cells, and that the non-regeneration sample has more state-3 mature, state-4 (transition 2) and state-5 (post-endosymbiotic) cells (Fig. 4f).

We further verified our hypothetical endosymbiotic cell states in the regeneration paradigm by pulse–chase experiments (Methods). After cutting, *Xenia* sp. stalks were pulsed with EdU at day 3 and day 4 of regeneration. EdU was washed out, corals were allowed to continue regenerating and samples were collected at days 7, 9, 11, 13, 15, 17 and 19 (Extended Data Fig. 7a). Using FACS (Extended Data Fig. 7b–h), we calculated the percentages of EdU$^+$ alga-containing cells out of all alga-containing *Xenia* cells, and the percentages of all alga-containing *Xenia* cells out of all *Xenia* cells. We found an increase of EdU$^+$ alga-containing *Xenia* cells up to regeneration day 13, which may account for the increase in uptake of algae during tentacle growth (as tentacles have more alga-containing cells than the stalk) (Extended Data Figs. 4d, 7i). Thereafter, the percentage of total alga-containing cells remained constant, but the percentage of EdU$^+$ alga-containing cells gradually decreased (Extended Data Fig. 7i, j). Thus, these results support our hypothesis that the endosymbiotic cells progress from a progenitor state through an alga-uptake state and a mature alga-containing state, followed by loss of their algae.

Analysis of differentially expressed genes suggests the roles of each state in endosymbiotic cell lineage development and function. For example, the state-1 pre-endosymbiotic cells express *WNT7b* and *WNT11*, which may regulate progenitor-cell proliferation and differentiation through the Wnt signalling pathway[31,32]. Among 24 genes

preferentially expressed in state 3, 13 are endosymbiotic markers that are expressed at higher levels in the FACS-isolated alga-containing *Xenia* cells than that in the alga-free cells. By contrast, none of the genes preferentially expressed in state 5 is an endosymbiotic marker. Instead, state-5 cells preferentially express several oxidative-stress-response genes (see Supplementary Table 6 for detailed descriptions). Because increased oxidative stress is observed upon cellular ageing and during coral bleaching[33–35], state-5 cells may represent a natural ageing state of endosymbiotic cells that are no longer able to hold on to their algae. Additional molecular studies exploring the function of the differentially expressed genes in each state are needed to further validate our five-state hypothesis.

Summary and outlook

Here we demonstrate the power of genomic and bioinformatic tools in studying coral biology. The *Xenia* sp. genome encodes essential components of RNA interference, such as Dicer and Ago, and DNA repair pathway proteins, which should enable the development of gene-manipulation tools to determine the mechanism of endosymbiosis. Although we focused on studying the endosymbiotic cell lineage, the regenerative processes for the other cell clusters can be similarly investigated in future analyses. Our studies suggest that *Xenia* endosymbiotic cells exist in five progressive states that are dynamic between homeostatic conditions and the regeneration process (Fig. 4f). It will be important to further understand the endosymbiotic lineage progression under different environmental stressors and to test whether efficient recovery from bleaching relies on state-1 pre-endosymbiotic cells. It is also feasible to test whether forced regeneration by fragmenting bleached corals can stimulate the expansion of state-1 pre-endosymbiotic cells and the restoration of endosymbiosis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2385-7.

1. Davy, S. K., Allemand, D. & Weis, V. M. Cell biology of cnidarian–dinoflagellate symbiosis. *Microbiol. Mol. Biol. Rev.* **76**, 229–261 (2012).
2. Putnam, H. M., Barott, K. L., Ainsworth, T. D. & Gates, R. D. The vulnerability and resilience of reef-building corals. *Curr. Biol.* **27**, R528–R540 (2017).
3. McFadden, C. S., Reynolds, A. M. & Janes, M. P. DNA barcoding of xeniid soft corals (Octocorallia: Alcyonacea: Xeniidae) from Indonesia: species richness and phylogenetic relationships. *Syst. Biodivers.* **12**, 247–257 (2014).
4. Sproles, A. E. et al. Phylogenetic characterization of transporter proteins in the cnidarian–dinoflagellate symbiosis. *Mol. Phylogenet. Evol.* **120**, 307–320 (2018).
5. Matthews, J. L. et al. Optimal nutrient exchange and immune responses operate in partner specificity in the cnidarian–dinoflagellate symbiosis. *Proc. Natl Acad. Sci. USA* **114**, 13194–13199 (2017).
6. Yuyama, I., Ishikawa, M., Nozawa, M., Yoshida, M. A. & Ikeo, K. Transcriptomic changes with increasing algal symbiont reveal the detailed process underlying establishment of coral–algal symbiosis. *Sci. Rep.* **8**, 16802 (2018).
7. Pinzón, J. H. et al. Whole transcriptome analysis reveals changes in expression of immune-related genes during and after bleaching in a reef-building coral. *R. Soc. Open Sci.* **2**, 140214 (2015).
8. Wolfowicz, I. et al. *Aiptasia* sp. larvae as a model to reveal mechanisms of symbiont selection in cnidarians. *Sci. Rep.* **6**, 32366 (2016).
9. Lehnert, E. M. et al. Extensive differences in gene expression between symbiotic and aposymbiotic cnidarians. *G3 (Bethesda)* **4**, 277–295 (2014).
10. Neubauer, E. F. et al. A diverse host thrombospondin-type-1 repeat protein repertoire promotes symbiont colonization during establishment of cnidarian–dinoflagellate symbiosis. *eLife* **6**, e24494 (2017).
11. Neubauer, E. F., Poole, A. Z., Weis, V. M. & Davy, S. K. The scavenger receptor repertoire in six cnidarian species and its putative role in cnidarian–dinoflagellate symbiosis. *PeerJ* **4**, e2692 (2016).
12. Wood-Charlson, E. M., Hollingsworth, L. L., Krupp, D. A. & Weis, V. M. Lectin/glycan interactions play a role in recognition in a coral/dinoflagellate symbiosis. *Cell. Microbiol.* **8**, 1985–1993 (2006).
13. Zheng, X. et al. Lamins organize the global three-dimensional genome from the nuclear periphery. *Mol. Cell* **71**, 802–815 (2018).
14. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
15. Kayal, E. et al. Phylogenomics provides a robust topology of the major cnidarian lineages and insights on the origins of key organismal traits. *BMC Evol. Biol.* **18**, 68–86 (2018).
16. Chen, G., Ning, B. & Shi, T. Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet.* **10**, 317 (2019).
17. Herring, C. A., Chen, B., McKinley, E. T. & Lau, K. S. Single-cell computational strategies for lineage reconstruction in tissue systems. *Cell. Mol. Gastroenterol. Hepatol.* **5**, 539–548 (2018).
18. Sebe-Pedros, A. et al. Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-seq. *Cell* **173**, 1520–1534 (2018).
19. Hwang, J. S. et al. Nematogalectin, a nematocyst protein with GlyXY and galectin domains, demonstrates nematocyte-specific alternative splicing in *Hydra*. *Proc. Natl Acad. Sci. USA* **107**, 18539–18544 (2010).
20. David, C. N. et al. Evolution of complex structures: minicollagens shape the cnidarian nematocyst. *Trends Genet.* **24**, 431–438 (2008).
21. Silverstein, R. L., Li, W., Park, Y. M. & Rahaman, S. O. Mechanisms of cell signaling by the scavenger receptor CD36: implications in atherosclerosis and thrombosis. *Trans. Am. Clin. Climatol. Assoc.* **121**, 206–220 (2010).
22. Kang, W. & Reid, K. B. DMBT1, a regulator of mucosal homeostasis through the linking of mucosal defense and regeneration? *FEBS Lett.* **540**, 21–25 (2003).
23. End, C. et al. DMBT1 functions as pattern-recognition molecule for poly-sulfated and poly-phosphorylated ligands. *Eur. J. Immunol.* **39**, 833–842 (2009).
24. Cenik, B., Sephton, C. F., Kutluk Cenik, B., Herz, J. & Yu, G. Progranulin: a proteolytically processed protein at the crossroads of inflammation and neurodegeneration. *J. Biol. Chem.* **287**, 32298–32306 (2012).
25. Popov, I. K., Ray, H. J., Skoglund, P., Keller, R. & Chang, C. The RhoGEF protein Plekhg5 regulates apical constriction of bottle cells during gastrulation. *Development* **145**, dev168922 (2018).
26. Meyer, E. & Weis, V. M. Study of cnidarian–algal symbiosis in the "omics" age. *Biol. Bull.* **223**, 44–65 (2012).
27. Barott, K. L., Venn, A. A., Perez, S. O., Tambutté, S. & Tresguerres, M. Coral host cells acidify symbiotic algal microenvironment to promote photosynthesis. *Proc. Natl Acad. Sci. USA* **112**, 607–612 (2015).
28. Alessandrini, F., Pezzè, L. & Ciribilli, Y. LAMPs: shedding light on cancer biology. *Semin. Oncol.* **44**, 239–253 (2017).
29. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
30. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
31. Afelik, S., Pool, B., Schmerr, M., Penton, C. & Jensen, J. Wnt7b is required for epithelial progenitor growth and operates during epithelial-to-mesenchymal signaling in pancreatic development. *Dev. Biol.* **399**, 204–217 (2015).
32. O'Brien, L. L. et al. Wnt11 directs nephron progenitor polarity and motile behavior ultimately determining nephron endowment. *eLife* **7**, e40392 (2018).
33. Downs, C. A. et al. Oxidative stress and seasonal coral bleaching. *Free Radic. Biol. Med.* **33**, 533–543 (2002).
34. Mydlarz, L. D. & Jacobs, R. S. An inducible release of reactive oxygen radicals in four species of gorgonian corals. *Mar. Freshwat. Behav. Physiol.* **39**, 143–152 (2006).
35. Finkel, T. & Holbrook, N. J. Oxidants, oxidative stress and the biology of ageing. *Nature* **408**, 239–247 (2000).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment other than the bioinformatic analyses.

Coral aquarium

The coral aquarium is established in a tank (Reefer 450 system, Red Sea). The artificial seawater, made from Coral Pro Salt (Red Sea), was first incubated with live rocks for two months before introducing *Xenia* sp., other corals, fish, snails and hermit crabs. The aquarium is maintained at about 80 °F with about 25% change of seawater every 1–2 weeks. The light is provided by Hydra 26 HD LED (Aqua Illumination) with 60% power on during 10:00 to 19:00. The fish were fed with fish pellets (New Life Spectrum Marine Fish Formula) and Green Marine Algae (Ocean Nutrition).

The *Xenia* sp. was obtained from a local coral aquarium shop called CTE Aquatics. We performed taxonomy analysis by amplifying *ITS2* rDNA region of Symbiodiniaceae species with primers (SYM_VAR_5.8S2, GAATTGCAGAACTCCGTGAACC and SYM_VAR_REV, CGGGTTCWCTTGT YTGACTTCATGC)[36]. Sequence analysis showed that the *Xenia* sp. in our aquarium contains multiple Symbiodiniaceae species, of the genus *Durusdinium*. In all of our experiments, samples of polyps or colonies were randomly selected from the aquarium. We selected polyps that appeared fully grown in size, and colonies that were easy to break off from their attachment sites. We will share our live *Xenia* sp. with any researchers upon request. We have also deposited some frozen and fixed coral colonies, along with genomic DNA and total RNA, at the Smithsonian National Museum of Natural History (catalogue no., USNM 1613385).

Genomic DNA isolation from *Xenia* sp.

To enable Nanopore DNA sequencing, we modified a protocol[37] that allowed the isolation of long DNA fragments. For each DNA preparation, one or two *Xenia* sp. colonies containing about 30 polyps were collected from the aquatic tank and washed 3 times for 5 min each with Ca^{2+}- and Mg^{2+}-free artificial seawater (449 mM NaCl, 9 mM KCl, 33 mM Na_2SO_4, 2.15 mM $NaHCO_3$, 10 mM Tris-HCl, 2.5 mM EGTA, pH 8.0). Tentacles were cut away, as they secrete a lot of mucus (which affected the quality of the isolated DNA). The remaining stalks and the bases of individual *Xenia* colonies were placed in 100 µl DNAzol (Invitrogen) in a 1.5-ml microcentrifuge tube. The tissues were cut into small pieces by a scissor to make fragment sizes of about 1/10th of the original size. These fragments were further minced by a small pestle made for 1.5-ml microcentrifuge tubes (Fisher Scientific, 12-141-364). Then, 900 µl DNAzol was added, followed by vortexing the sample and then transferred to a 15-ml conical tube. Four millilitres of DNAzol and 50 µl of 10 mg/ml RNase A were then added to the tube and mixed, followed by incubation at 37 °C for 10 min. Then, 25 µl of 20 mg/ml proteinase K was then added, mixed and the tube was incubated at 37 °C for another 10 min. The sample was centrifuged at 5,000*g* for 10 min. The supernatant was transferred to another 15-ml tube. After adding 2.5 ml ethanol, the tube was gently mixed by inverting several times. The tube was left to stand at room temperature for 3 min followed by centrifugation at 1,000*g* for 10 min to pellet the DNA. The supernatant was discarded and the DNA pellet was resuspended in 500 µl TE (10 mM Tris-HCl, 1 mM disodium EDTA, pH 8.0). After the DNA had dissolved, 500 µl of phenol:chloroform:isoamyl alcohol (25:24:1) was added, and the tube was placed on the Intelli-Mixer RM-2S for mixing using programme C1 at 35 rpm for 10 min. The mixture was then transferred to a 2-ml phase-lock gel (QuantaBio, Cat. 2302820) and centrifuged at 4,500 rpm for 10 min. The aqueous phase was transferred into a new 2-ml tube, 200 µl 5 M ammonium acetate and 1.5 ml ice-cold ethanol were added followed by centrifugation at 10,000*g* for 10 min to pellet DNA. The pellet was washed twice with 1 ml 80% ethanol. After removing as much ethanol as possible, the DNA pellet was left to dry at 42 °C for 1 min, and then resuspended in 50 µl TE buffer.

Illumina sequencing

Genomic DNA prepared as in 'Genomic DNA isolation from *Xenia* sp.' was fragmented into about 400 bp, and libraries were made with ThruPLEX DNA-Seq kit (TaKaRa) according to the manufacturer's manual. These libraries were sequenced using the NEXseq500 platform with NextSeq 500/550 High Output Reagent Cartridge v2 (Illumina).

Nanopore sequencing

Genomic DNA was used to build Nanopore sequencing libraries with Ligation Sequencing Kit (SQK-LSK108, Oxford Nanopore Technologies), following the manufacturer's manual. For the first three runs, genomic DNA was not fragmented, to generate long reads. To obtain more reads, for the fourth run of Nanopore sequencing, genomic DNA was sheared to 8–10 kilobases by g-TUBE (Covaris, 520079). The libraries were sequenced in R9.4.1 flow cells on a MinION device (Oxford Nanopore Technologies). MinKNOW (v.1.7.3) was used to collect raw signal and Albacore (v.2.3.3) was used for base-calling. All the data were combined for genome assembly.

Hi-C

To perform Hi-C on *Xenia* sp. tissue, we modified a previously published protocol for nuclear in situ ligation[13], as described in detail.

Fix and dissociate tissues (step 1). Eight polyps (about 10^8 cells) were fixed with 4% paraformaldehyde (PFA) overnight. After washing twice with 3.3× PBS[38] and dissociating the tissue in 2 ml 3.3× PBS using a 7-ml glass Dounce tissue grinder (Wheaton), another 3 ml 3.3× PBS was added. The mixture was then transferred to a 15-ml conical tube and centrifuged at 1,000*g* for 3 min (Sorvall Lynx 6000 centrifuge, ThermoFisher Scientific). The pellet was washed once with 5 ml 3.3× PBS.

Nuclear permeabilization and chromatin digestion (step 2). The pellet from step 1 was resuspended in 10 ml ice-cold Hi-C lysis buffer (10 mM Tris, pH 8.0, 10 mM NaCl, 0.2% NP-40, 1× protease inhibitors cocktail (Roche, 04693132001)) and rotated for 30 min at 4 °C followed by centrifugation at 1,000*g* for 5 min at 4 °C. The pellet was resuspended with 1 ml ice-cold 1.2× NEB3.1 (120 µl NEB3.1 to 880 µl ddH_2O) buffer and transferred to a 1.5-ml microcentrifuge tube followed by centrifugation at 1,000*g* for 5 min at 4 °C. The pellet was washed again with 1 ml ice-cold 1.2× NEB3.1 followed by centrifugation. After removing the supernatant, 400 µl 1.2× NEB3.1 buffer and 12 µl of 10% SDS were added to the pellet. P200 pipette tip was used to thoroughly resuspend and dissociate the pellet. The mixture was then incubated at 65 °C for 10 min at 950 rpm in a Thermomixer (Eppendorf). After cooling the mix on ice for 5 min, 40 µl 20% Triton X-100 was added to the mixture to neutralize the SDS. After carefully mixing by pipetting with a P200 pipette tip and inverting the tube several times, the mixture was then incubated at 37 °C for 60 min with rotation (950 rpm) in a Thermomixer. To digest the crosslinked genomic DNA, 30 µl of 50 U/µl BglII (NEB R0144M) was added to the mixture and incubated overnight at 37 °C with rotation at 950 rpm in a Thermomixer.

Fill in 5′ overhang generated by BglII digestion with biotin (step 3). A nucleotide mix containing dATP, dGTP and dTTP was made by adding 1 µl each of 100 mM dATP, dGTP and dTTP into 27 µl ddH_2O. To the 480.0 µl BglII-digested nuclear preparation from the above step 2, 4.5 µl of the nucleotide mix, 15 µl 1 mM biotin-16-dCTP (Axxora, JBS-NU-809-BIO16) and 10 µl 5 U/µl Klenow (NEB, M0210L) were added followed by incubation at 37 °C for 90 min with intermittent gentle shaking at 700 rpm for 10 s after every 20 s using Thermomixer. The tube was also taken out and inverted every 15–20 min. After this incubation, the mixture was kept on ice.

Proximity ligation (step 4). The mixture from step 3 was transferred to a 50-ml conical tube followed by adding 750 µl 10× T4 ligase buffer (NEB B0202S, no PEG), 75 µl 100× BSA (NEB), 6,140 µl water, 25 µl 30 U/µl T4 DNA ligase (Thermo Scientific, EL0013), and incubating at 16 °C overnight.

Reverse crosslink and DNA isolation (step 5). To the reaction mixture from step 4, 25 µl of 20 mg/ml proteinase K (Invitrogen, 25530-049) was added and the mixture was divided equally into 8×1.5-ml microcentrifuge tubes (about 950 µl per tube). The tubes were then incubated overnight at 65 °C with rotation at 950 rpm in a Thermomixer. The next day, 3 µl 20 mg/ml proteinase K was added to each tube followed by incubation at 65 °C for 2 h with mixing in Thermomixer. The mixtures were combined into one 50-ml conical tube. After cooling down to room temperature, 10 ml phenol (pH 8.0) (Sigma) was added and mixed by vortex for 2 min. The mixture was then centrifuged for 10 min at 3,000*g* (Sorvall Lynx 6000 centrifuge). The supernatant containing the DNA was mixed with 10 ml phenol:chloroform (1:1) (pre-warmed to room temperature) and vortexed for 2 min. The whole mixture was then transferred to a 50-ml MaXtract High Density tube (Qiagen, 129073) and centrifuged at 1,500*g* for 5 min (Sorvall Lynx 6000 centrifuge). The top phase containing the Hi-C DNA was transferred to a 50-ml conical tube and the volume (usually about 10 ml) was adjusted to 10 ml with 1× TE as needed. To pellet the DNA, 1 ml 3 M Na-acetate, 5 µl 15 mg/ml GlycoBlue (Invitrogen AM9515) and 10 ml isopropanol were added to the mixture and incubated at −80 °C for >1 h. The DNA was then pelleted by centrifugation at 17,000*g* for 45 min at 4 °C (Sorvall Lynx 6000 centrifuge). The Hi-C DNA pellet was resuspended in 450 µl 1× TE and transferred to a 1.5-ml microcentrifuge tube followed by adding 500 µl phenol:chloroform (1:1). After mixing by vortex, the mix was centrifuged at 18,000*g* for 5 min at room temperature. The top aqueous layer was collected into another tube followed by adding 40 µl 3M Na-acetate, 1 µl 15 mg/ml GlycoBlue (Invitrogen AM9515, 300 µl) and 1 ml ice-cold 100% ethanol. After incubating at −80 °C for >30 min, the DNA was centrifuged at 21,000*g* for 30 min at 4 °C. The DNA pellet was washed with freshly prepared 70% ethanol and air-dried, followed by dissolving in 45 µl EB (10mM Tris, pH 8.0). The contaminated RNA in the DNA preparation was digested by adding 0.5 µl 10 mg/ml RNaseA and incubated at 37 °C for 30 min.

Remove biotin from the free DNA (unligated DNA) ends (step 6). To remove the biotin at the free DNA ends, 1.0 µl 10 mg/ml BSA (NEB, 100×), 10 µl 10× NEB 2.1 buffer, 1 µl 10 mM dATP, 1 µl 10 mM dGTP and 5 µl T4 DNA polymerase (NEB M0203S), and 42 µl water were added to 40 µl (about 3 µg) Hi-C DNA preparation from step 5. The mixture was divided into two equal aliquots in 2 PCR tubes and incubated at 20 °C for 4 h. Then, 2 µl of 0.5 M EDTA was added to each of the two tubes to stop the reaction. The Hi-C DNA was then purified using the Clean and Concentrator Kit (ZYMO, D4013) followed by elution with 50 µl EB.

Biotin pull-down of DNA and second DNA digestion (step 7). In brief, 60 µl of Dynabeads MyOne Streptavidin C1 (Invitrogen) was washed in 1.5-ml non-sticking microcentrifuge tubes (Ambion) with 200 µl 2× binding buffer (10 mM Tris, pH 8, 0,1 mM EDTA, 2 M NaCl) twice, followed by resuspension in 50 µl 2× binding buffer. The 50 µl Hi-C DNA from step 6 was added followed by rotating for 30 min using Intelli-Mixer (ELMI) at room temperature. The beads were collected using a magnetic stand and washed with 100 µl 1× binding buffer followed by washing with 100 µl 1× NEB4 buffer twice and resuspending in 50 µl 1× NEB4 buffer. The DNA on beads was digested using 1 µl 10 U/µl AluI (NEB, R0137S) at 37 °C for 60 min. The beads were collected on a magnetic stand followed by washing with 100 µl 1× binding buffer, and then 100 µl EB. The beads were resuspended in 30 µl EB.

A-tailing (step 8). The 30-µl beads from step 7 were mixed with 5 µl NEB Buffer 2, 10 µl 1 mM dATP, 2 µl H_2O, 3 µl Klenow (3′–5′ exo-) (NEB M0212L) and incubated at 37 °C for 45 min. After the reaction, the beads were collected by a magnetic stand followed by washing with 100 µl 1× binding buffer and then 100 µl EB. The beads were resuspended in 50 µl EB.

Sequencing adaptor ligation (step 9). The 50-µl beads from step 8 was mixed with 3.75 µl sequencing adaptor (TruSeq RNA Sample Prep Kit v.2), 10 µl 1× T4 DNA ligase buffer, 3 µl T4 DNA Ligase (30 U/µl) (Thermo Scientific, EL0013) and incubated at room temperature for 2 h. The beads were collected by a magnetic stand followed by washing twice with 400 µl 1× binding buffer + 0.05% Tween, 200 µl 1× binding buffer, and then 100 µl EB. The beads were resuspended in 40 µl EB. To release the DNA from the beads, the mixture was incubated at 98 °C for 10 min and then centrifuged at 500 rpm to pellet the streptavidin beads.

Sequencing library preparation (step 10). TruSeq RNA Library Prep Kit was used to make DNA sequencing library (eight PCR cycles were used) and the DNA was sequenced by NextSeq 500.

scRNA-seq

For each of the six scRNA-seq library preparation, 1 polyp, 8 tentacles, and 2 stalks or 2 regenerating stalks of *Xenia* sp. were dissociated into single cells in 1 ml digestion buffer, containing 3.6 mg/ml dispase II (Sigma, D4693), 0.25 mg/ml liberase (Sigma, 5401119001), 4% L-cysteine in Ca^{2+}-free seawater (393.1 mM NaCl, 10.2 mM KCl, 15.7 mM $MgSO_4{\cdot}7H_2O$, 51.4 mM $MgCl_2{\cdot}6H_2O$, 21.1 mM Na_2SO_4, and 3 mM $NaHCO_3$, pH 8.5) and incubated for 1 h at room temperature. After digestion, fetal bovine serum was added to a final concentration of 8% to stop enzymatic digestion. The cell suspension was filtered through a 40-µm cell strainer (FALCON). A low concentration (0.1 µg/ml) of DAPI that can only be taken up by dead cells was used to measure cell viability. Only cell suspensions in which more than 90% of cells that did not take up DAPI were used. Cells were counted by haemocytometer and diluted with the same 4% L-cysteine in Ca^{2+}-free seawater used in the digestion buffer into 1,000 cells per µl. Around 17,000 cells per sample were used for single-cell library preparation using the 10x Genomics platform with Chromium Single Cell 3′ Library and Gel Bead Kit v.2 (PN-120267) (v.2 chemistry) or Chromium Next GEM Single Cell 3′ GEM, Library and Gel Bead Kit v.3.1 (PN-1000121, v.3 chemistry), Single Cell 3′ A Chip Kit (PN-1000009) or Chromium Next GEM Chip G Single Cell Kit (PN-1000127), and i7 Multiplex Kit (PN-120262). For the scRNA-seq library construction, we followed the 10x protocol exactly. In brief, for v.2 chemistry, 17.4 µl cell suspension and 16.4 µl nuclease-free water were mixed with 66.2 µl reverse transcription master mix. Of this 100 µl mix, 90 µl was loaded into the chip provided in the Single Cell 3′ A Chip Kit. For v.3 chemistry, 16.5 µl cell suspension and 26.7 µl nuclease-free water were mixed with 31.8 µl reverse transcription master mix. Of this 75 µl mix, 70 µl was loaded into the Chromium Next GEM Chip G. After barcoding, cDNA was purified and amplified with 11 PCR cycles. The amplified cDNA was further purified and subjected to fragmentation, end repair, A-tailing, adaptor ligation and 14 cycles of sample index PCR. Libraries were sequenced using Illumina NextSeq 500 for paired-end reads. Read 1 is 26 bp (v.2 chemistry) or 28 bp (v.3.1 chemistry) and read 2 is 98 bp.

In our initial scRNA-seq using the 10x Genomics v.2 chemistry, we obtained fewer unique molecular identifiers (UMIs) (median number, 801) and genes (median number, 467) per *Xenia* cell compared to other model organisms, such as the mouse thymus[39] (median UMI 5,802 and median gene number 2,178), but higher than in *Nematostella*[18] (median UMI 541 and median gene number 278). The new and improved v.3 chemistry substantially improved our scRNA-seq. We captured more cells per library (v.3 7,874 versus v.2 2,883), a higher number of median genes per cell, (v.3 943 versus v.2 467) and median UMI per cell (v.3 2,027 versus v.2 801). Our v.3 dataset has lower quality than those of

the mouse thymus[39] and *Hydra*[40] scRNA-seq datasets (Supplementary Table 1). This suggests that, even using v.3 chemistry, the presence of seawater and/or *Xenia*-sp.-specific features may contribute to the reduced scRNA-seq quality.

We noticed the mapping rate in v.3 chemistry is lower than in v.2 chemistry. We sequenced more reads for the v.3 libraries, because v.3 captured more total cells and more RNA molecules per cell. Although we sequenced more for the v.3 libraries, we obtained lower sequence saturation (on average, 79.6% in v.3 libraries and 92.6% in v.2 libraries). Because the v.2 and v.3 reagent contents are proprietary information, it is difficult for us to assess why the two methods gave different results. Regarding our library preparation, the v.3 method entailed 22% of the total volume coming from the cell suspension in the Ca^{2+}-free seawater, while in the v.2 method, 17.4% of the total volume came from the Ca^{2+}-free seawater cell suspension. We therefore know that one difference between the two methods is that the salt concertation in v.3 library preparation is higher than that in the v.2 library preparation. The higher salt concentration in v.3 could lead to a higher RNA extraction efficiency in the v.3 library preparation, which could contribute to the difference between our v,2- and v,3-based scRNA-seq. Althought the 10x platform worked well for the *Xenia* sp. we studied here, it is important to keep in mind that modifications may be needed for successful scRNA-seq for other marine cnidarians.

Quantification of endosymbiotic *Xenia* cells by microscopy and FACS

To quantify the endosymbiotic cell percentage in *Xenia*, we first applied a microscopy-based strategy. By imaging cryo-preserved tissue sections stained with 1 µg/ml DAPI that labelled all nuclei, we determined the total number of *Xenia* cells per section by counting the number of *Xenia* cell nuclei: these nuclei are easily differentiated from the alga nuclei when overlapped with the autofluorescence signal in far red channel from algae. The number of *Xenia* cells containing alga is estimated by counting the number of algae surrounded by *Xenia* tissue. The estimated percentage by this method is on average 2–6%, depending on whether the sections were taken from stalks or tentacles (Extended Data Fig. 4d). The limitation of this method is that some algae that appear to be inside the tissue may be between *Xenia* cells and not inside cells. Therefore, this estimate could represent an upper limit of the percentage of alga-containing *Xenia* cells.

In the second method, we used FACS to separate free algae and algae contained inside the *Xenia* cells. *Xenia* polyps were dissociated into single-cell suspension with the same preparation method as described in 'scRNA-seq'. The cells were fixed with 1% (final concentration) formaldehyde on ice for 1 h, followed by 0.2% Triton X-100 permeablization and 1 µg/ml DAPI staining. We first separated free algae and alga-containing *Xenia* cells according to the algae autofluorescence in the Cy5.5 channel. Free algae and algae inside *Xenia* cells should have different forward scatter (FSC) and side scatter (SSC) signals because the alga inside *Xenia* cells is enclosed by the *Xenia* cellular membrane structure. Thus, we used FSC and SSC to further gate the total population of algae into two subpopulations. Microscopy analyses showed that this gating separated free algae and alga-containing *Xenia* cells. To determine the total *Xenia* cell number, *Xenia* cells together with algal cells were gated according to DAPI-positive signal followed by gating with the Cy5.5 signal. The total *Xenia* cells were calculated as alga-free *Xenia* cells plus the alga-containing *Xenia* cells. On the basis of these FACS analyses, we were able to estimate the percentage of alga-containing *Xenia* cells in *Xenia* polyps to be about 2% of total *Xenia* cells. The illustration of this FACS sorting can be found in Extended Data Fig. 7b–g. Because the procedure of single-cell dissociation may cause an alga-containing *Xenia* cell to lose its alga, the approximately 2% of alga-containing *Xenia* cells obtained by the FACS method probably represents an underestimation. Thus, we estimate the fraction of alga-containing *Xenia* cells to be about 2–6%.

Bulk RNA-seq

Total RNA was isolated from 3 polyps, 32 tentacles or 6 stalks by RNeasy Plus Mini Kit (Qiagen). To obtain additional transcriptomes from different cell types, we dissociated coral tissue into individual cells according to a previously published method[41] and subjected the dissociated cells to OptiPrep-based cell separation[42]. Cells with different densities were separated into four layers, and RNA was isolated from each layer with RNeasy Plus Mini Kit (Qiagen). For transcriptome of FACS-isolated alga-containing and alga-free cells, three polyps were dissociated with the same protocol as used in the scRNA-seq and the dissociated cells were subjected to FACS. Cy5.5-positive and -negative cells were collected as alga-containing and alga-free cells, respectively, and used for total RNA extraction as above. cDNA libraries were built according to TruSeq Stranded mRNA Library Prep Kit (Illumina) and subjected to Illumina NextSeq 500 for sequencing. For gene annotation, paired-end sequencing of 75 bp for each end was used. For FACS-isolated bulk-cell transcriptomes, single-end sequencing of 75 bp was used.

Xenia regeneration, BrdU labelling and EdU pulse–chase

Individual *Xenia* sp. polyps were placed into a well of 24-well cell-culture plate (Corning) containing 1 ml artificial seawater from our aquatic tank. The polyps were allowed to settle in the well for 5–7 days before cutting away the tentacles. After cutting, there were a lot algae released into the seawater, which together with the free algae living inside the cavity of the coral could serve as alga reservoirs for the uptake of algae during regeneration.

For the BrdU labelling experiments, 0.5 mg/ml BrdU was added into the well 2 d before sample collection. The BrdU-labelled stalks were fixed by 4% PFA overnight, followed by washing with PBST (PBS+0.1% Tween 20) twice for 10 min each. The stalk was then balanced with 30% sucrose overnight followed by embedding in OCT, frozen in dry ice bathed in ethanol and subjected to cryo-sectioning. The slides were washed with PBS 3 times for 5 min each time followed by treating with 2 M HCl containing 0.5% Triton X-100 for 30 min at room temperature. The slides were then incubated with PBST (0.2% Triton X-100 in PBS) 5 min for 3 times each followed by blocking with 10% goat serum and then incubating with mouse anti-BrdU antibody (ZYMED, 18-0103, 1:200 dilution in 10% goat serum) at 4 °C overnight. Slides were washed with PBST 3 times for 10 min each followed by incubation with the secondary antibody (Invitrogen) for 1 h at room temperature and washing with PBST 3 times for 10 min each. The nuclei were counterstained with Hoechst 33342 and the signal was visualized using a confocal microscope (Leica). Clear BrdU signal in the nucleus labelled by Hoechst was counted as a $BrdU^+$ cell. If the *Xenia* $BrdU^+$ nucleus was juxtaposed to an alga, it was counted as an alga-containing $BrdU^+$ *Xenia* cell.

For EdU pulse–chasing experiments, the regenerating *Xenia* stalks were incubated with 1 mM EdU during regeneration day 3 and day 4. After washing out EdU, the coral was incubated with artificial seawater and samples were collected on regenerating days 7, 9, 11, 13, 15, 17 and 19. The samples were dissociated into single-cell suspensions followed by fixing with 1% formaldehyde at 4 °C overnight as described in 'scRNA-seq'. The fixed cells were pelleted at 800*g* for 5 min and further fixed with 4% PFA for two days to block the autofluorescence in the 488-nm channel. Then, the EdU click chemistry was carried out using the Click-iT EdU Cell Proliferation Kit (Invitrogen, C10337) according to manufacturer's protocol. The cells were further stained with DAPI, and then analysed by FACS as described in Extended Data Fig. 7 and 'Quantification of endosymbiotic *Xenia* cells by microscopy and FACS'.

Whole-mount RNA ISH

To perform RNA ISH on *Xenia*, we modified the whole-mount RNA ISH protocol for zebrafish[43].

For making gene-specific sense or anti-sense probes, we designed primers (Supplementary Table 7) to genes of interest for PCR to amplify

gene fragments from *Xenia* sp. cDNA. The T3 promoter sequence was added to the 5′ of the reverse primers so that the PCR products could be directly used for synthesizing anti-sense RNA probes by T3 RNA polymerase (Promega, P2083) using DIG RNA Labelling Mix (Roche, 11277073910). DIG-labelled RNA probes were purified by RNA Clean and Concentrator-5 (ZYMO), heated to 80 °C for 10 min, immediately transferred on ice for 1 min, and then diluted in Prehyb$^+$ buffer (50% formamide, 5× saline–sodium citrate buffer (SSC, 0.75M NaCl, 0.075M sodium citrate), 50 µg/ml heparin, 2.5% Tween 20, 50 µg/ml single-stranded DNA (Sigma, D1626)) to a final concentration of 0.5 µg/ml, and stored at −20 °C until use.

Xenia polyps were relaxed in Ca^{2+}-free seawater for 30 min and then fixed in 4% PFA in Ca^{2+}-free seawater overnight at 4 °C. Fixed polyps were washed with PBST (0.1% Tween 20 in PBS) twice for 10 min each, and then incubated in 100% methanol at −20 °C overnight.The next day, the tissues were washed sequentially in 75%, 50% and 25% methanol for 5 min each and then washed in PBST for 10 min. They were then treated with 50 µg/ml proteinase K in PBST for 20 min followed by a brief wash in PBST. The tissues were post-fixed in 4% PFA at room temperature for 20 min and then washed with PBST 2 times for 10 min each. Prehybridization was performed in Prehyb$^+$ at 68 °C for 2 h, followed by incubation with probes in Prehyb$^+$ overnight at 68 °C. To probe gastrodermis markers, 2% SDS (final concentration) was added to help the probes to penetrate the tissue. After probes were removed, samples were washed sequentially in 2× SSC (0.3 M NaCl and 0.03 M sodium citrate) containing 50% formamide for 20 min twice, 2× SSC containing 25% formamide for 20 min, 2× SSC for 20 min twice, and 0.2× SSC for 30 min 3 times each, all at 68 °C. Then, samples were washed in PBST at room temperature for 10 min and incubated in DIG blocking buffer (1% ISH blocking reagent (Roche, 11096176001) in maleic acid buffer (0.1 M maleic acid, 0.15 M NaCl, pH 7.5) for 1 h at room temperature, followed by incubation in anti-DIG antibody (anti-digoxigenin-AP (Roche, 11093274910)) at 1:5,000 dilution in DIG blocking buffer overnight at 4 °C. The next day, the samples were washed in PBST for 10 min 3 times each at room temperature, then in 9.5T buffer (100 mM Tris-HCl pH 9.5, 50 mM $MgCl_2$, 100 mM NaCl, 0.1% Tween 20) for 10 min 3 times each at room temperature. Hybridization signals were revealed by incubation in BCIP/NBT buffer (1 SIGMAFAST BCIP/NBT tablet (Sigma, B5655) in 10 ml H_2O)) at 4 °C until brown–purplish colours were sufficiently dark. For this study, the colour development took 48 h. The samples were then washed in PBST twice for 10 min each. The samples were post-fixed in 4% PFA overnight at 4 °C, followed by washing in PBST twice for 10 min each, and then washed in methanol for 3 h at room temperature. The tissues were kept in PBS and imaged using SMZ1500 microscope (Nikon) under Ring Light System (Fibre-Lite). For cross- sections of stalks, the whole-mount sample was processed for cryo-section as described in '*Xenia* regeneration, BrdU labelling and EdU pulse–chase'.

RNAscope ISH assay for *LePin* and *Granulin 1* expression

To visualize RNA expression in endosymbiotic cells, we used the ultra-sensitive RNAscope ISH approach (Advanced Cell Diagnostics (ACD)). *LePin- or Granulin-1*-specific oligonucleotide probes were ordered from ACD (see Supplementary Table 7 for further information). The fluorescent RNAscope assay was carried out by RNAscope Multiplex Fluorescent Reagent Kit v.2 (ACD) according to the manufacturer's protocol. The chromogenic assay was carried out by RNAscope 2.5 HD Duplex Detection Kit (ACD), according to manufacturer's protocol. Both assays used the cryo-section of the fixed *Xenia* polyp prepared according to the manufacturer's protocol.

Genome assembly

Sequencing data from Nanopore were used to initiate the genome assemble by Canu (v.1.7)[44]. The assembled genome was further polished with Illumina short reads by Nanopolish (v.0.9.2, https://github.com/jts/nanopolish) with 5 cycles, which resulted in 1,482 high-quality contigs for the diploid genome. The diploid genome assembly was separated into haploid by HaploMerger2[45]. The haploid genome assembly was further subject to Hi-C assisted scaffolds by 3D de novo assembly pipeline, Juicer (v.1.5)[14]. By aligning all the Illumina genomic sequencing data with the assembled genome, we found 0.45% single nucleotide polymorphism (SNP) within the whole assembled genome of the *Xenia* sp.

Gene annotation

The funannotate genome annotation pipeline (v.1.3.3, https://github.com/nextgenusfs/funannotate) was used to annotate the *Xenia* sp. genome. In brief, transcriptome data were assembled by Trinity (v.2.6.6)[46] and used to generate the gene models based on the presence of mRNA by PASApipeline (v.2.3.2)[47]. These gene models were used as training sets to perform de novo gene prediction by AUGUSTUS (v.3.2.3)[48] and GeneMark-ES Suite (v.4.32)[49]. All gene models predicted by PASApipeline, AUGUSTUS and GeneMark were combined and subjected to EVidenceModeller to generate combined gene models[50]. The predicted genes were filtered out if more than 90% of the sequence overlapped with repeat elements as identified by RepeatMasker and RepeatModeler (http://www.repeatmasker.org). PASA was further used to add 3′ and 5′ untranslated region sequences to the remaining predicted genes. Pfam (v.31.0), Interpro (v.67.0), Uniprot (v.2018_03), BUSCO (v.1.0)[51] databases and eggnog-mapper (v.1.3)[52] were used to annotate the function of these gene models. Among all the predicted genes, 23,939 (82.5%) gene models were supported by transcriptome data because they have detectable reads (reads number >0). Among these models, 20,397 have read numbers >5.

Phylogeny tree analysis

We used OrthoFinder (v.2.2.7) to find orthologues from different species on the basis of protein sequences from 13 species listed Fig. 1d, and inferred the species tree[53,54]. In brief, 'orthofinder -S diamond -t 22 -M msa -f fasta_files' was used to generate the result. Diamond (v.0.9.21) was used for sequence search and OrthoFinder grouped 308,348 genes (83.8% of total) into 19,244 orthogroups. One thousand six hundred and one orthogroups, according to previously reported method[55], with a minimum 10 species having single-copy genes, were used to infer the species tree. These orthogroups were subjected to multiple sequence alignment by MAFFT (v.7.407) and columns with more than eight gaps were trimmed. The trimmed alignment with 73.6% data occupancy (see Source Data for Fig. 1d) was used to infer the maximum likelihood unrooted species tree by FastTree (v.2.1.10) with the default configuration in OrthoFinder. This species tree was further rooted by the STRIDE algorithm, which has been demonstrated to correctly root the species tree spanning a wide range of time scales and taxonomic groups[56].

Single-cell clustering and marker gene identification

The raw single-cell sequencing data were de-multiplexed and converted to FASTAQ format by Illumina bcl2fastaq (v.2.20.0) software. Cell Ranger (v.3.1.0, https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome) was used to de-multiplex samples, process barcodes and count gene expression. The sequence was aligned to the annotated *Xenia* sp. genome and only the confidently mapped and non-PCR duplicated reads were used to generate gene expression matrix for each library with 'cellranger count' command. The expression matrix of Cell-Ranger-identified cells from each library was read into R and further analysed with Seurat (v.3.0.2)[57]. Cells with UMI numbers less than 400 or mitochondria gene expression >0.2% of total reads were excluded for downstream analysis. To further remove outliers, we calculated the UMI number distribution detected per cell and removed cells in the top 1% quantile.

To remove batch effect and integrate data from different libraries, we applied the Seurat v.3 method for data integration[57]. For each dataset, we identified the top 1,000 genes with the highest dispersion. We used the top 1,000 genes in the non-regeneration sample as anchor features to identify anchors between different non-regeneration datasets. The first 20 dimensions were used to generate the integrated data. Dimensional reduction was carried out on the integrated data, and used for further clustering analysis. Clustering and marker gene identification in non-regeneration condition was further performed with Seurat v.3. The cell clusters in regeneration samples were identified with the label transfer method in Seurat v.3. All violin plots were generated using Seurat VlnPlot function.

Identification of *Xenia* sp. cells performing endosymbiosis with Symbiodiniaceae

The bulk transcriptome data of FACS-isolated alga-containing or alga-free *Xenia* cells were aligned to *Xenia* sp. genome by STAR (v.2.5.3a)[58]. Individual gene expression (reads per kilobase of transcript, per million mapped reads) for each sample were calculated by RSEM (v.1.3.0)[59]. The gene-expression levels of each bulk RNA-seq of FACS-isolated cells were compared with the gene-expression levels calculated using average UMI number for each gene in each cell cluster identified by scRNA-seq. The Pearson correlation coefficient was calculated for each comparison.

Pseudotime analysis

To infer the trajectory of endosymbiotic *Xenia* cells, we integrated scRNA-seq data of regenerating and non-regenerating samples using Seurat v.3. All cells belonging to the endosymbiotic cell cluster (cluster 16, total of 382 cells) were subjected to Monocle (v.2.10.1)[29] analyses. To find the variable genes among these cells for downstream analysis, we grouped these cells into three subclusters with Monocle clusterCells function (with default setting for most parameters, except for num_clusters = 4, which generated 3 clusters). Each of these three subclusters contains 247, 53 or 82 cells. The top 1,000 differentially expressed genes between these three subclusters were used as ordering genes to construct the trajectory by DDRTree algorithm. The differentially expressed genes along pseudotime were detected using the differentialGeneTest function in Monocle. The cell numbers in each of the five predicted endosymbiotic cell states are state 1 = 36, state 2 = 109, state 3 = 155, state 4 = 45 and state 5 = 37.

RNA velocity

RNA velocity estimation was carried out using the velocyto.R program (http://velocyto.org, v.0.6), according to the instructions[30]. In brief, velocyto used raw data of the regeneration sample to count the spliced (mRNA) and unspliced intron reads for each gene to generate a .loom file. This .loom file was loaded into R (v.3.6.1) using the read.loom.matrices function and used to generate the RNA velocity map. The RNA velocity map was projected into the *t*-SNE space that was identified by Seurat.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

We have uploaded all raw genomic, bulk RNA-seq and scRNA-seq data to NCBI (BioProject PRJNA548325). The genome files are available at http://cmo.carnegiescience.edu/data; we have also made the genome data interactive using UCSC genome browser, http://genome.ucsc.edu/cgi-bin/hgTracks?hubUrl=http://cmo.carnegiescience.edu/gb/hub.txt&genome=xenSp1. We allow anyone interested to explore the predicted proteomes of *Xenia* and 14 other cnidarian using our blast server: http://c-moor.carnegiescience.edu:4567. All scRNA-seq analyses and results are available at GitHub: https://github.com/ciwemb/endosymbiosis. Select intermediate RDS objects are available at: http://cmo.carnegiescience.edu/data. We have worked to prototype a web portal to organize all the above links. This work-in-progress has a goal of making research findings, experimental protocols and computational data available to the scientific community. As the portal involves information beyond this study, we are still working with colleagues to best design it so that it will be easy to use and informative. The portal can be accessed at: http://cmo.carnegiescience.edu. Source Data are provided with this paper.

Code availability

R Markdown codes are available at https://github.com/ciwemb/endosymbiosis. For convenience, processed data and code can be downloaded with the following Unix commands: git clone https://github.com/ciwemb/endosymbiosis; wget -r -np -nH --reject = "index.html*" http://cmo.carnegiescience.edu/endosymbiosis.

36. Hume, B. C. C. et al. An improved primer set and amplification protocol with increased specificity and sensitivity targeting the *Symbiodinium* ITS2 region. *PeerJ* **6**, e4816 (2018).
37. Urban, J. M., Bliss, J., Lawrence, C. E. & Gerbi, S. A. Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. Preprint at https://www.biorxiv.org/content/10.1101/019281v3 (2015).
38. Rosental, B., Kozhekbaeva, Z., Fernhoff, N., Tsai, J. M. & Traylor-Knowles, N. Coral cell separation and isolation by fluorescence-activated cell sorting (FACS). *BMC Cell Biol.* **18**, 30 (2017).
39. Yue, S., Zheng, X. & Zheng, Y. Cell-type-specific role of lamin-B1 in thymus development and its inflammation-driven reduction in thymus aging. *Aging Cell* **18**, e12952 (2019).
40. Siebert, S. et al. Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution. *Science* **365**, eaav9314 (2019).
41. Helman, Y. et al. Extracellular matrix production and calcium carbonate precipitation by coral cells in vitro. *Proc. Natl Acad. Sci. USA* **105**, 54–58 (2008).
42. Mass, T. et al. Cloning and characterization of four novel coral acid-rich proteins that precipitate carbonates in vitro. *Curr. Biol.* **23**, 1126–1131 (2013).
43. Hu, M. et al. Liver-enriched gene 1, a glycosylated secretory protein, binds to FGFR and mediates an anti-stress pathway to protect liver development in zebrafish. *PLoS Genet.* **12**, e1005881 (2016).
44. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
45. Huang, S. et al. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res.* **22**, 1581–1588 (2012).
46. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
47. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
48. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
49. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).
50. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
51. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
52. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
53. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
54. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
55. Emms, D. M. & Kelly, S. STAG: species tree inference from all genes. Preprint at https://www.biorxiv.org/content/10.1101/267914v1 (2018).
56. Emms, D. M. & Kelly, S. STRIDE: species tree root inference from gene duplication events. *Mol. Biol. Evol.* **34**, 3267–3278 (2017).
57. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
58. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
59. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

Acknowledgements We thank F. Tan and A. Pinder for assistance with all the sequencing; F. Tan and Q. Zhang for assistance in establishing the Carnegie Coral and Marine Organisms web portal and GitHub; Y. Bai for assistance with cell sorting; M. Sepanski for assistance with electron microscopy; N. Marvi for the coral sketch; and L. Hugendubler and M. Watts for maintaining the coral aquarium. This work was supported by Gordon and Betty Moore

Foundation, Aquatic Symbiosis no. GBMF9198 (https://doi.org/10.37807/GBMF9198, Y.Z.), NIH/NIGMS GM106023 (Y.Z.), GM110151 (Y.Z.), NIH/NIAMS AR060042 (C.-M.F.) and AR071976 (C.-M.F.).

Author contributions C.-M.F. and Y.Z. conceived and supervised the project. M.H., C.-M.F. and Y.Z. designed experiments. M.H. performed the experiments. M.H. and X.Z. analysed the data. M.H., X.Z., C.-M.F. and X.Z. interpreted the data and wrote the manuscript.

Competing interests The authors declare no competing interests.

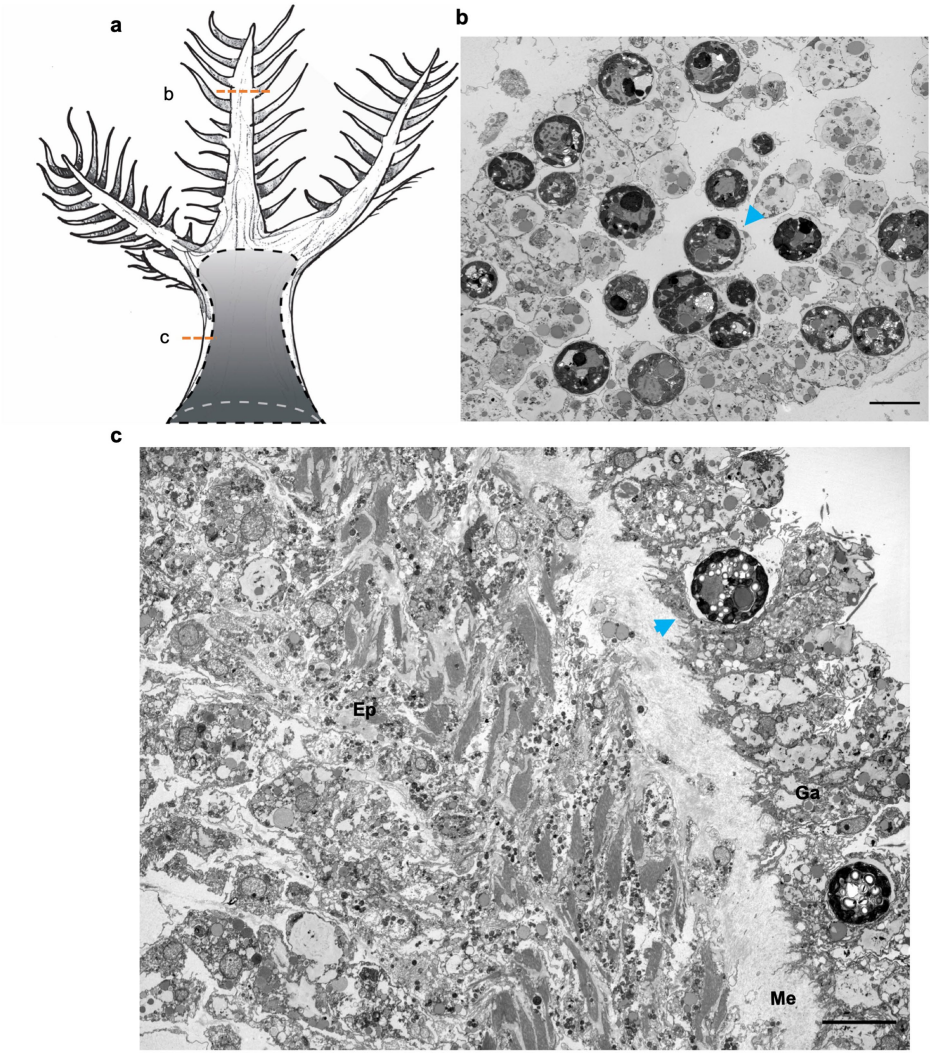
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2385-7.

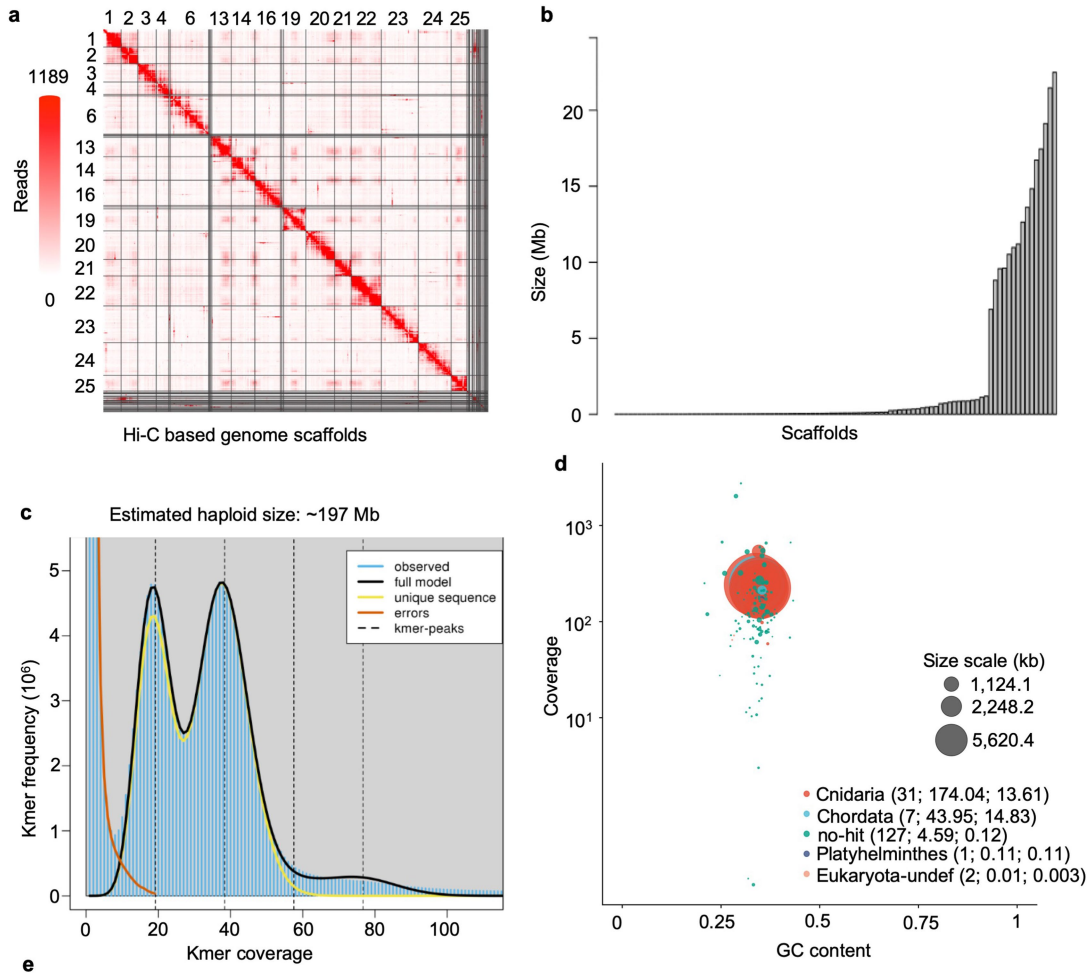
Correspondence and requests for materials should be addressed to M.H., C.-M.F. or Y.Z.

Peer review information *Nature* thanks Mónica Medina Munoz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.



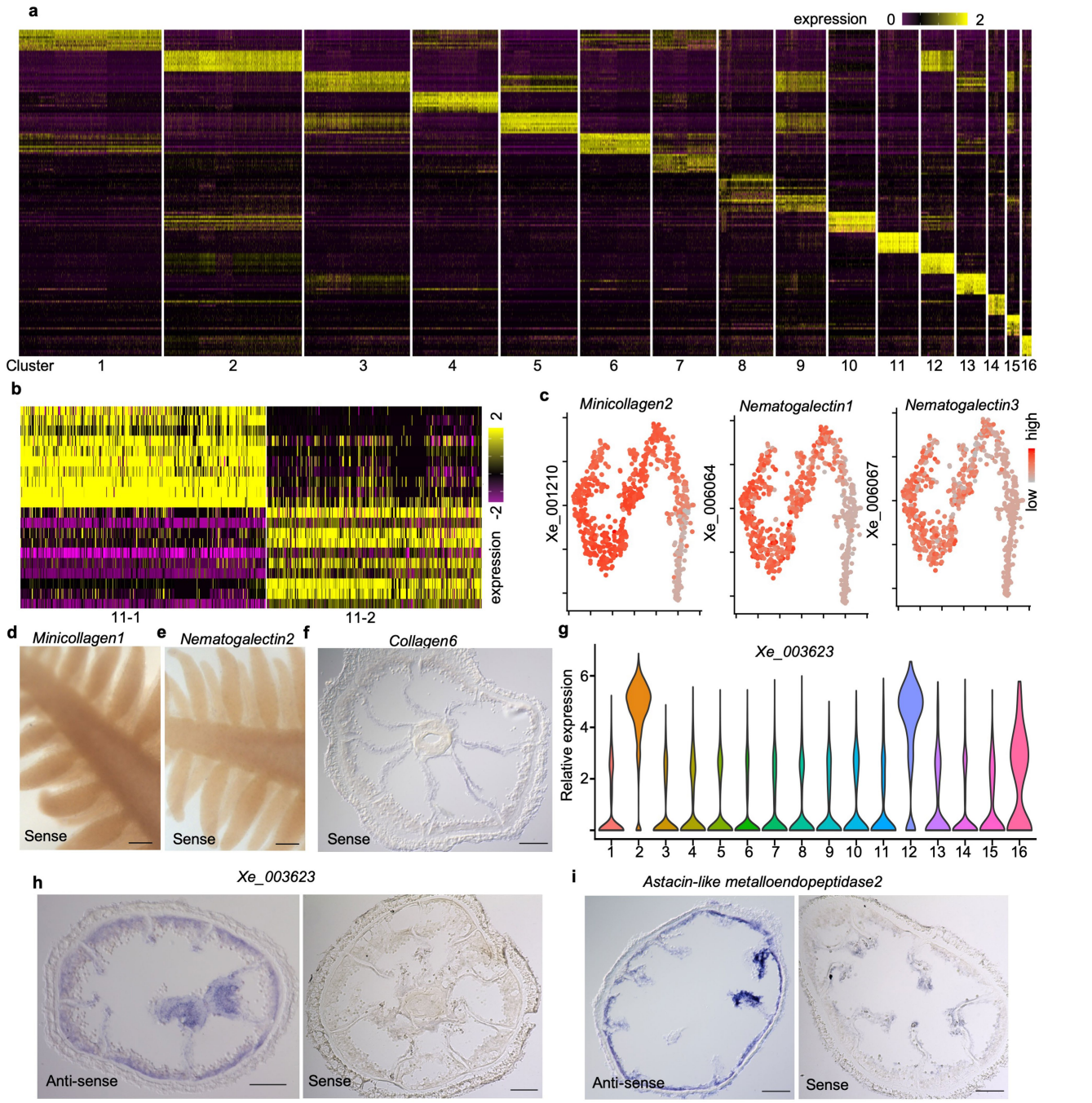
Extended Data Fig. 1 | Electron microscopy analysis of *Xenia* sp. **a**, Illustration of a *Xenia* polyp. The orange dashed lines indicate where the electron microscopy images were taken, shown in **b** and **c**. **b**, **c**, Electron microscopy images. Ep, epidermis; Ga, gastrodermis; Me, mesoglea. Blue arrowheads, alga-containing *Xenia* cells. Five independent polyps from two independent experiments were used for electron microscopy. Scale bars, 10 µm.



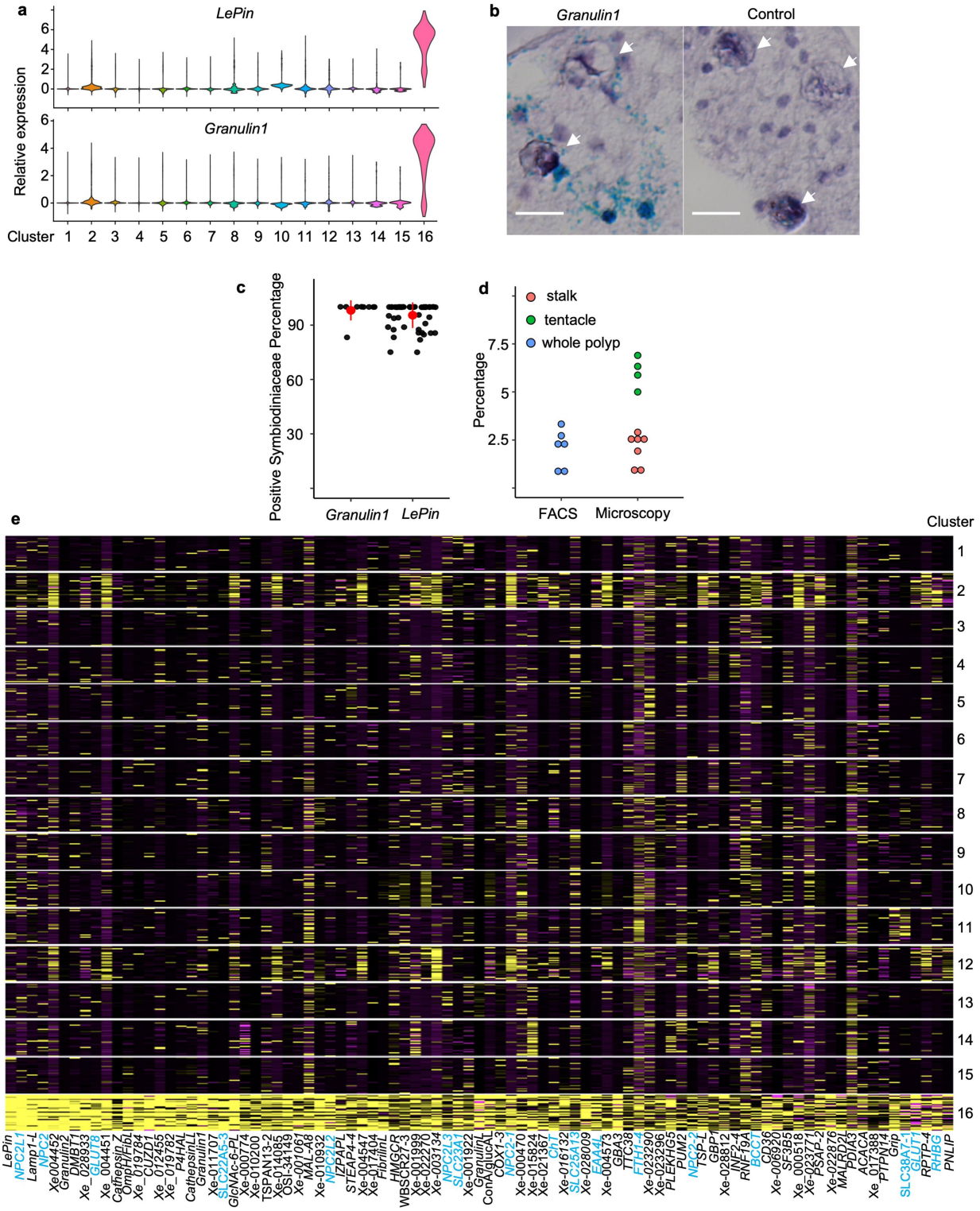
e

Parameters	Values
Genome size, predicted by GenomeScope	197 Mb
Total size of genome assembly, in 168 scaffolds	222,699,500 bp
Total contig size, in 556 contigs	222,505,550 bp
Scaffold N50	14,832,246 bp
Longest scaffold	22,481,500 bp
Contig N50	1,122,448 bp
Longest contig	4,187,899 bp
Number of gene models	29015
Number of protein coding genes	28011
Number of apparently complete gene models*	27640
Number of predicted proteins with recognizable (E-value ≤1e-5) similarity in ncbi nr database	21783
Mean predicted protein length	414 amino acid
Repeat content	46.22%

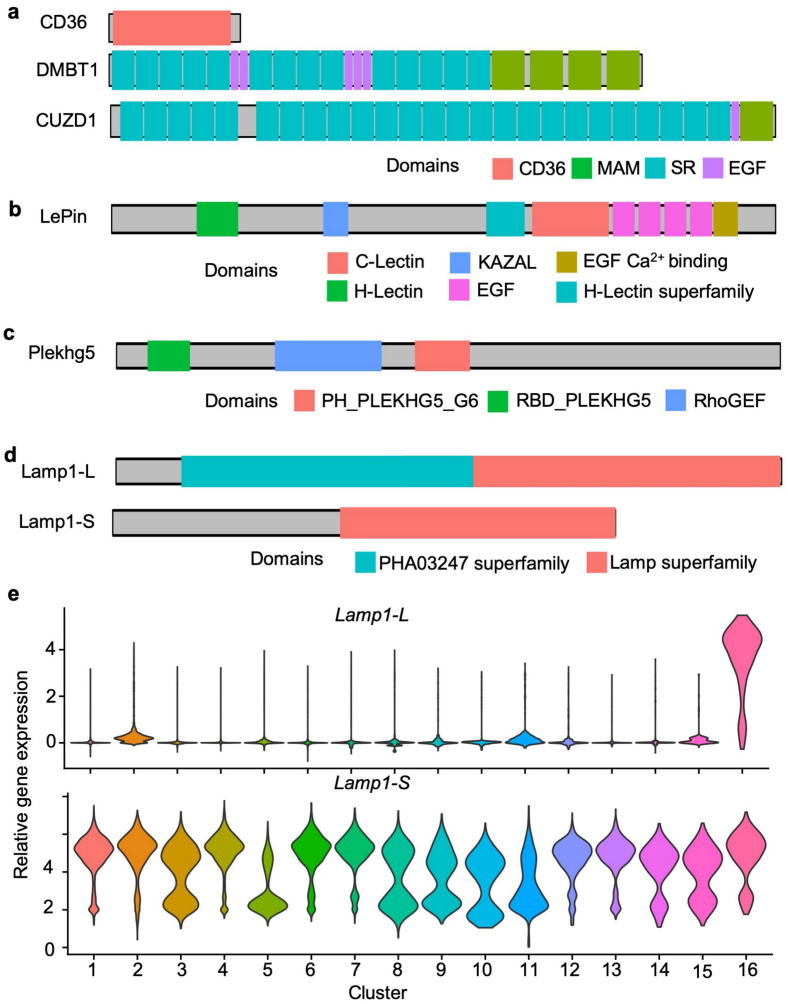
Extended Data Fig. 2 | Additional genome assembly data. a, Hi-C-based *Xenia* sp. genome assembly. The scaffolds are separated by grids demarcated by black lines. The numbers for the 15 longest scaffolds out of the total 168 scaffolds are shown. **b**, Size distribution of *Xenia* sp. genome scaffolds. Each bar on the *x* axis represents a scaffold. **c**, *Xenia* sp. genome is predicted to be diploid, as expected, with a haploid genome size of about 197 Mb, on the basis of GenomeScope analysis of Illumina short reads. **d**, Contamination analysis by BlobTools revealed a similar GC content and genomic coverage across most scaffolds. Each coloured circle in the graph represents a scaffold. Larger circles have longer scaffold lengths; the three grey circles provide the length scale used in the plot. The colour codes represent the closest species group that has the highest sequence similarities to the *Xenia* sp. scaffolds (the first number in each set of parentheses shows the *Xenia* scaffold number followed by the combined length of the scaffolds and scaffold N50 value (minimum contig length needed to cover 50% of the combined scaffold length) in Mb). **e**, A summary of *Xenia* sp. genome assembly and gene annotation. *Genes encoding protein sequences with apparent in frame start and stop codons.



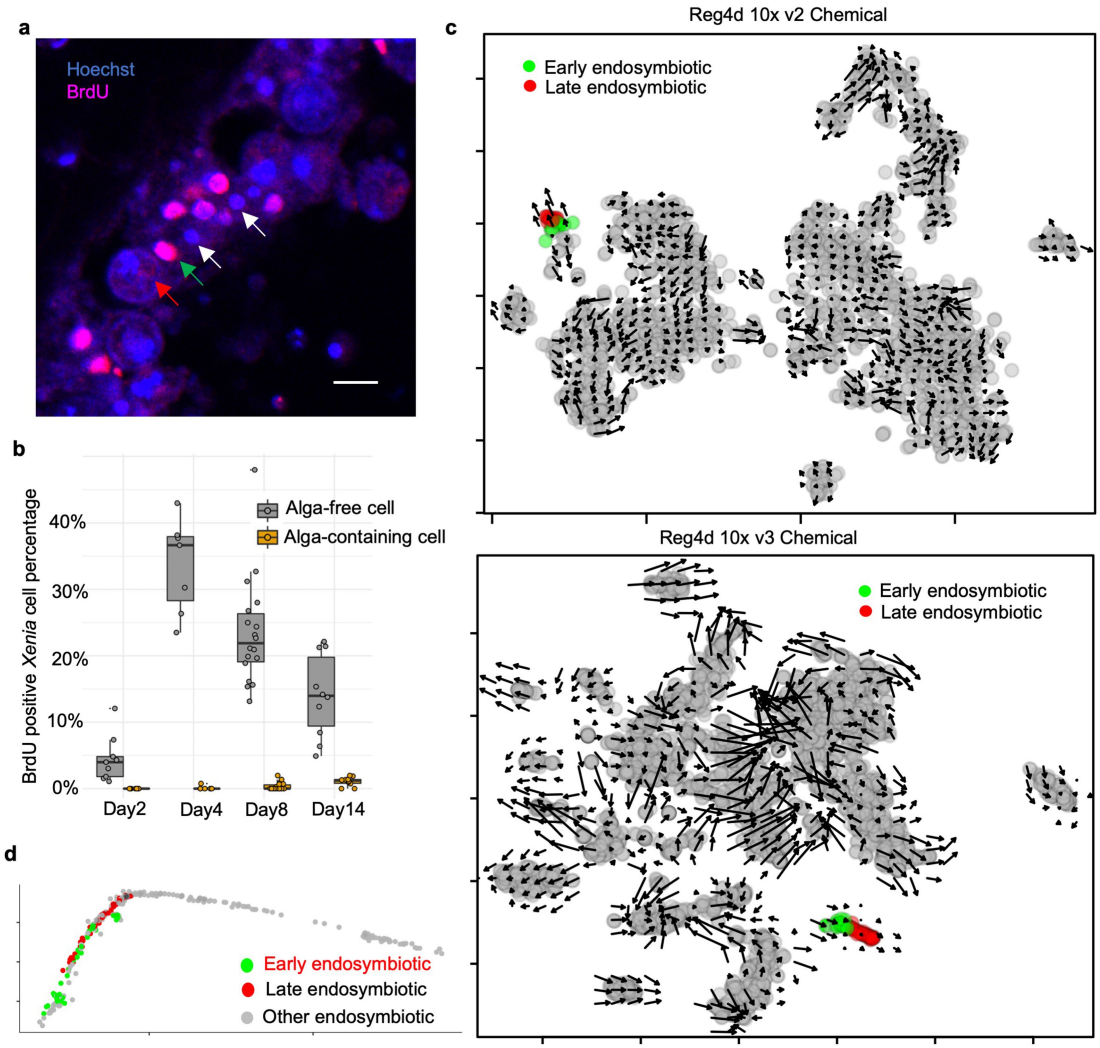
Extended Data Fig. 3 | Additional scRNA-seq analyses. a, Heat map showing differential gene expression patterns of all cells in the 16 assigned cell clusters (indicated at the bottom). Each column is one cell cluster, and each row represents one gene. **b**, Heat map showing differential gene expression patterns of two subclusters in cluster 11 (11-1 and 11-2). **c**, Expression levels (as in the coloured expression scale) of 3 cluster-11 markers, *Minicollagen 2*, *Nematogalectin 1* and *Nematogalectin 3*, are shown in the *t*-SNE plots. $n = 797$ cells. **d**–**f**, RNA ISH control with sense probe for *Minicollagen 1* (**d**), *Nematogalectin 2* (**e**) and *Collagen 6* (**f**). **g**, Expression levels of *Xe_003623*, a non-conserved and uncharacterized cluster-2 and -12 marker gene, in each of the 16 cell types defined by scRNA-seq. Violin plot (Methods) show the distribution of gene expression in each of the 16 clusters. Cell numbers in cell clusters 1–16 were 2,794; 2,704; 2,073; 1,679; 1,511; 1,374; 1,248; 1,069; 986; 923; 797; 649; 575; 321; 246; and 185, respectively. **h**, **i**, RNA ISH of *Xe_003623* (**h**) and *Astacin-like metalloendopeptidase 2* (**i**) using anti-sense and sense probes. In **d**–**f**, **h**, **i**, more than 12 polyps from 3 independent experiments were used for each probe. Scale bars, 100 µm.



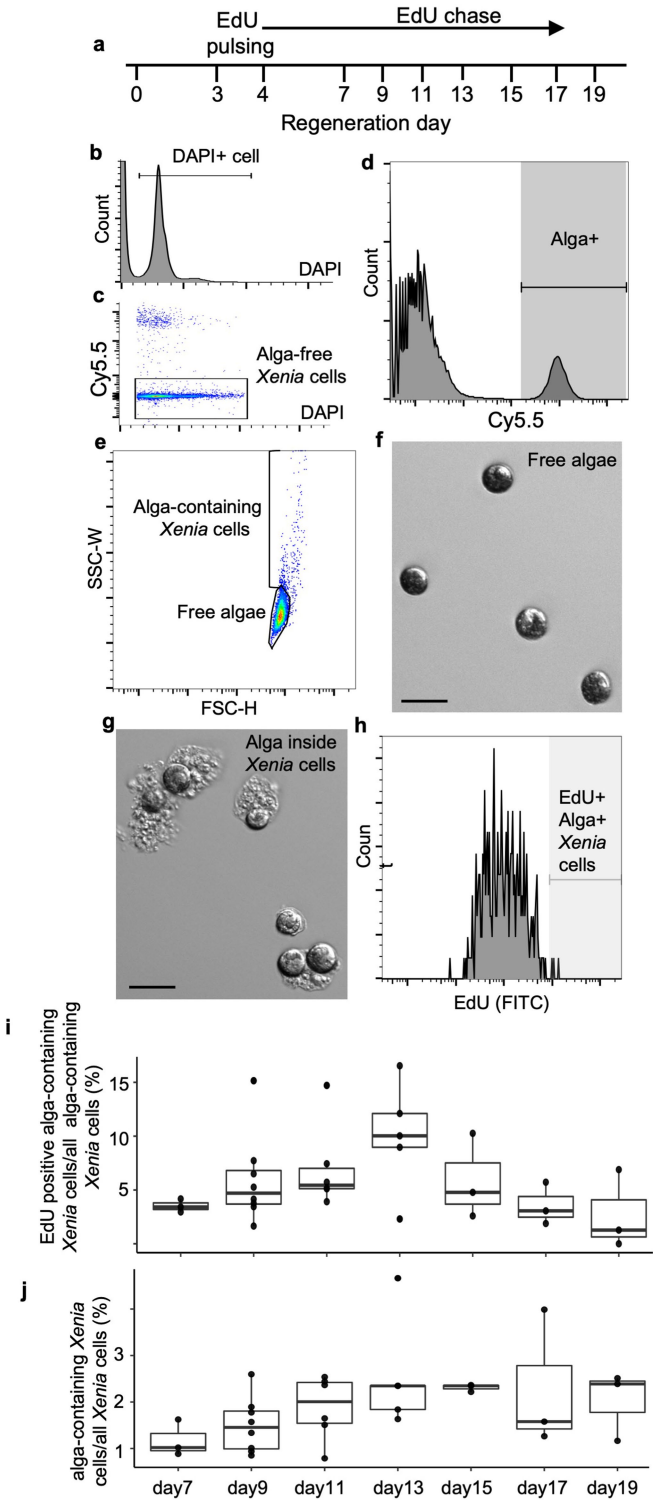
Extended Data Fig. 4 | Additional analyses for endosymbiotic cells. a, Violin plots of the expression profiles of *LePin* and *Granulin 1* in the 16 clusters defined by scRNA-seq. Violin plots show the distribution of gene expression in each of the 16 clusters. Cell numbers in cell clusters 1–16 are 2,794; 2,704; 2,073; 1,679; 1,511; 1,374; 1,248; 1,069; 986; 923; 797; 649; 575; 321; 246 and 185, respectively. **b**, Ultra-sensitive chromogenic RNA ISH by RNAscope probing for *Granulin 1* (left) and control (right). Positive signals are blue. Nuclei were counterstained as purple with haematoxylin. White arrows indicate algae of the Symbiodiniaceae. Six polyps from three independent experiments were used for each probe. Scale bars, 10 μm. **c**, Percentage of alga-containing cells with positive *Granulin 1* or *LePin* signal. Each black dot stands for one section. Red dots and lines stand for mean and s.d., respectively. Six polyps from three independent experiments were used for one gene or control. **d**, Percentage of alga-containing cells measured by FACS and microscopy. For FACS, each dot stands for an individual polyp. Three independent experiments were performed with each experiment, using two polyps. For microscopy, each dot stands for a section analysed in three polyps. **e**, Heat map showing the enrichment levels of the 89 marker genes in cluster 16 among all 16 cell clusters. Transporters are highlighted in blue.



Extended Data Fig. 5 | Selected endosymbiotic markers with known domains. a, The scavenger receptors (SR) CD36, DMBT1 and CUZD1. **b**, LePin. **c**, Plekhg5. **d**, Lamp1-L and Lamp1-S. **e**, Violin plots of expression profiles of *Lamp1-L* and *Lamp1-S* in 16 clusters defined by scRNA-seq. Violin plots show the gene-expression distribution in each cluster. Cell numbers in cell clusters 1–16 are 2,794; 2,704; 2,073; 1,679; 1,511; 1,374; 1,248; 1,069; 986; 923; 797; 649; 575; 321; 246 and 185, respectively.



Extended Data Fig. 6 | Additional analyses of endosymbiotic cell lineage. **a**, A representative image of BrdU labelling (pink), overlaid with Hoechst (blue DNA stain) in a cross-section of a regenerating *Xenia* sp. stalk. White, red and green arrows indicate BrdU-negative ($BrdU^-$) *Xenia* nuclei, an alga and a BrdU-positive ($BrdU^+$) *Xenia* nucleus juxtaposed to the alga, respectively. Three regenerating stalks were used in two independent experiments. **b**, Box plot. Percentages (*y* axis) of $BrdU^+$ *Xenia* cells at the indicated regeneration time points (*x* axis). Each dot represents data from one section. Three regenerating stalks from two independent experiments were pooled and plotted for each time point. About 7 to 18 sections were used for each group. The alga-containing proliferated *Xenia* cells were estimated as those with $BrdU^+$ nuclei juxtaposed to algae. The medians are indicated as lines in the box; the upper and lower edge of the boxes represent the upper and lower quartiles, respectively. **c**, Velocyto analysis of the scRNA-seq data from day-4 regenerating *Xenia* sp. stalks. Each dot represents a cell, and arrows indicate the directions of RNA velocity. The green and red endosymbiotic cell clusters were predicted as early and late-cell states, respectively, on the basis of the directions of the arrows. **d**, The distribution of early (green) and late (red) endosymbiotic cells predicted by velocyto in **c** on the pseudotime plot of all scRNA-seq data shows the start and the direction of progression of the endosymbiotic cell lineage.



Extended Data Fig. 7 | EdU pulse–chase analysis of endosymbiotic cells. **a**, Pulse–chase experiments. The regeneration stalk was labelled with EdU at regeneration day 3 and day 4. After washing out EdU, the samples were cultured, collected and analysed at the indicated days during chasing. **b**, Dissociated cells were processed by Click-iT to visualize EdU and stained with DAPI to label nuclei. Cells were sorting on the basis of DAPI. **c**, DAPI-positive cells were further sorted on the basis of Cy5.5 to estimate the number of the total alga-free *Xenia* cells. **d**, **e**, To estimate the number of alga-containing *Xenia* cells, free algae and alga-containing *Xenia* cells (alga$^+$ population) were first separated from all the other *Xenia* cells on the basis of the Cy5.5 signal (**d**). The alga$^+$ population was further separated into alga-containing *Xenia* cells and free algae based on the SSC and FSC signals (**e**). **f**, **g**, Microscopy confirmation of free algae (**f**) and alga-containing *Xenia* cells (**g**) sorted in **e**. Scale bars, 20 μm. In **c**–**h**, four independent experiments were carried out. **h**, The number of EdU-positive and alga-containing *Xenia* cells were further estimated on the basis of their strong EdU signal. **i**, Box plot of the percentage of EdU-positive and alga-containing *Xenia* cells among all alga-containing *Xenia* cells at the indicated days of chase. **j**, Box plot of the percentage of all alga-containing *Xenia* cells among all *Xenia* cells at the indicated days of chase. Each dot in **i**, **j** stands for one regenerating sample. Day 7, $n = 3$ polyps; day 9, $n = 8$ polyps; day 11, $n = 6$ polyps; day 13, $n = 5$ polyps; day 15, $n = 3$ polyps; day 17, $n = 3$ polyps; day 19, $n = 3$ polyps from 2 independent experiments were assayed. The medians are indicated as lines in the boxes; the upper and lower edges of the boxes represent the upper and lower quartiles, respectively.

Extended Data Table 1 | Summary of sequencing libraries for genome assembly

a

library	read number (M*)	read length	pair-end	data (G†)
1	27.2	150	No	4.09
2	19.7	150	No	2.91
3	73.9	75	No	5.54
4	127.4	150	Yes	38.2

b

library	read number	max (bp*)	mean (bp)	median (bp)	> 5 kb†	>10 kb	>20 kb	data (G‡)
run1	819,205	62,624	4,488	4,341	38.4%	3.5%	0.19%	3.68
run2	328,045	266,931	11,716	4,343	46.1%	29.9%	17.3%	3.84
run3	210,985	310,074	12,677	5,400	52.4%	32.7%	18.3%	2.67
run4	1,912,621	143,219	5,228	4,959	49.5%	7.27%	0.37%	10

a, Summary of Illumina sequencing for genome assembly. The table shows sequence information from four library preparations from four *Xenia* colonies for Illumina sequencing. Read number indicates the total number of reads obtained. Read length indicates the individual read length, by paired-end sequencing. Data indicate total sequence data in gigabases. M*, million; G†, gigabase.

b. Summary of Nanopore sequencing for genome assembly. Statistics of all sequence information from four different runs from four *Xenia* colonies of Nanopore sequencing, including maximum, mean and median read-length statistics, and the percentages of reads that have bigger sizes than the indicated number: >5 kb, >10 kb, and >20 kb. bp*, base pair; kb†, kilobase; G‡, gigabase.

Extended Data Table 2 | Transcriptomes for gene annotation

samples	library	read number (M*)	data (G†)
whole polyp	pair-end	34.2	5.12
Stalk	pair-end	36.0	5.40
Tentacle	pair-end	39.8	5.97
Regeneration 4d stalk	pair-end	38.9	5.84
Opti-Prep lv1	pair-end	29.6	4.43
Opti-Prep lv2	pair-end	31.8	4.76
Opti-Prep lv3	pair-end	33.6	5.03
Opti-Prep lv4	pair-end	31.1	4.68

A summary of all the transcriptome data used for gene annotation. RNA isolated from different samples as indicated were used for Illumina sequencing to cover as many expressed genes as possible. Opti-Prep, density-based separation of dissociated *Xenia* cells into four different layers (Methods). lv1, lv2, lv3, and lv4 indicate layer 1, layer 2, layer 3 and layer 4 cells, respectively (used to make the RNA-seq libraries). M*, million; G†, gigabase.

Extended Data Table 3 | Comparisons of *Xenia* sp. genome assembly with the assembled genomes of the indicated and published cnidarians

Parameter \ species	*Xenia sp.*	*Exaiptasia diaphana*	*Nematostella vectensis*	*Acropora digitifera*
Predicted genome size (Mb)	197	260	329	420
Assembly size (Mb)	222.7	258	356	419
Total contig size (Mb)	222.5	213	297	365
Total contig size as % of assembly size	99.9	82.5	83.4	87
Contig N50 (Kb)	1,122	14.9	19.8	10.9
Scaffold N50 (Kb)	148,322	440	472	191
Number of gene models	29015	29269	27273	23668
Number of complete gene models	27640	26658	13343	16434
Mean exon length (bp)	204	354	208	230
Mean intron length (bp)	448	638	800	952
Mean protein length (number of amino acids)	414	517	331	424
Predicted protein BUSCO (n=978) completeness	90.1%	89.4%	93.8%	54.8%

The number of gene models indicates the predicted gene model number, whereas the number of complete gene models represents the number of genes with clearly predicted in-frame start and stop codons. Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness was assessed by conserved gene models in metazoans using BUSCO3. The *Exaiptasia diaphana* gene model v.1.0 was downloaded from http://aiptasia.reefgenomics.org/download/aiptasia_genome.proteins.fa.gz. The *Acropora digitifera* gene model v.0.9 was downloaded from https://marinegenomics.oist.jp/coral/download/adi_aug101220_pasa_gene.fa.gz. The *Nematostella vectensis* gene model was downloaded from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000001593_45351.fasta.gz.

Corresponding author(s): Minjie Hu, Chen-Ming Fan, Yixian Zheng

Last updated by author(s): Apr 12, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	MinKNOW (v1.7.3) , Albacore (v2.3.3), bcl2fastaq (v2.20.0)
Data analysis	Canu (v1.7), Nanopolish (v0.9.2), HaploMerger2, Juicer (v1.5), Funannotate (v1.3.3),Trinity (v2.6.6), PASApipeline (v2.3.2), AUGUSTUS (v3.2.3), GeneMark-ES Suite (v4.32), eggnog-mapper (v1.3), OrthoFinder (v2.2.7), Diamond (v0.9.21), MAFFT (v7.407), FastTree (v2.1.10), Cell Ranger (v3.1.0), Seurat (v3.0.2), STAR (v2.5.3a), RSEM (v1.3.0), Monocle (v2.10.1), velocyto.R (v0.6), R(v3.6)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequence data for this study is available in NCBI BioProject under accession PRJNA548325. Assembled genome and gene annotation are available at http://cmo.carnegiescience.edu/data. The scRNA analysis code is available at https://github.com/ciwemb/endosymbiosis

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. We followed standards in the biology field.
Data exclusions	For single cell RNA-seq analysis, based on pre-established criteria for single-cells, in order to remove empty droplet, or droplet with potential dead cells or potential doublets, cells with UMI numbers less than 400 or mitochondria gene expression >0.2% were filtered out. To further remove outliers, we calculated the UMI number distribution detected per cell and removed cells in the top 1% quartile.
Replication	Each experiment was replicated with multiple independent animals. To draw a conclusion, at least two independent experiments were carried. All replicates were successful.
Randomization	Xenia colonies or polyps were randomly chosen from the aquarium tank
Blinding	Quantification of LePin signal was blinded by de-identifying samples.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	mouse anti-BrdU antibody, from ZYMED. The catlog number is 18-0103, ZBU30 clone, Lot Number 00460071R. The dilution is 1:200.
Validation	The BrdU antibody was validated by a lot of studies listed in the manufactory's website: https://www.thermofisher.com/antibody/product/BrdU-Antibody-clone-ZBU30-Monoclonal/03-3900. It has been applied in IF, IHC, FACS in Chemical, Chicken, Mouse, Rabbit and Rat. We validated it by the lack of staining when BrdU was not added into the sample.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Xenia sp. was cultured in laboratory aquarium tank. We can not yet tell their age and sex.
Wild animals	The Xenia sp. used in this study was originally from the wild, but we obtained it from an aquarium shop in Baltimore.
Field-collected samples	The study didn't involve samples collected from field
Ethics oversight	The study of Xenia or some other cnidaria does not yet have ethical oversight.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Xenia polyps were dissociated into single cell suspension with the same method for single cell RNA-seq. More details are provided in the method.
Instrument	BD FACSAria™ III
Software	BD FACSDiva Software v6.1.3
Cell population abundance	All Xenia cells were divided into two population, algea-containing and algea-free, based on Cy5.5 signal. The two population have distinct Cy5.5 signal and are easy to separate. There's almost no contamination as confirmed by microscopy inspection on the sorted population. The EdU positive algea-containing population is a small population and the percentage is plotted in Extended fig 7i
Gating strategy	Detailed gating strategy is described in the method.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

Hidden neural states underlie canary song syntax

https://doi.org/10.1038/s41586-020-2397-3

Received: 24 February 2019

Accepted: 26 March 2020

Published online: 17 June 2020

Check for updates

Yarden Cohen[1]✉, Jun Shen[2], Dawit Semu[1], Daniel P. Leman[1], William A. Liberti III[1,3], L. Nathan Perkins[1], Derek C. Liberti[4,5,6], Darrell N. Kotton[4,5,6] & Timothy J. Gardner[1,7]✉

Coordinated skills such as speech or dance involve sequences of actions that follow syntactic rules in which transitions between elements depend on the identities and order of past actions. Canary songs consist of repeated syllables called phrases, and the ordering of these phrases follows long-range rules[1] in which the choice of what to sing depends on the song structure many seconds prior. The neural substrates that support these long-range correlations are unknown. Here, using miniature head-mounted microscopes and cell-type-specific genetic tools, we observed neural activity in the premotor nucleus HVC[2–4] as canaries explored various phrase sequences in their repertoire. We identified neurons that encode past transitions, extending over four phrases and spanning up to four seconds and forty syllables. These neurons preferentially encode past actions rather than future actions, can reflect more than one song history, and are active mostly during the rare phrases that involve history-dependent transitions in song. These findings demonstrate that the dynamics of HVC include 'hidden states' that are not reflected in ongoing behaviour but rather carry information about prior actions. These states provide a possible substrate for the control of syntax transitions governed by long-range rules.

Canary songs, like many flexible behaviours, contain complex transitions—points at which the next action depends on memory for choices made several steps in the past. Songs are composed of syllables produced in trilled repetitions known as phrases (Fig. 1a) that are about 1 s long and are sung in sequences, typically 20–40 s long. The order of phrases in a song exhibits long-range syntax rules[1]. Specifically, phrase transitions following about 15% of the phrase types depend on the preceding sequence of 2–5 phrases. These long-range correlations extend over dozens of syllables, spanning time intervals of several seconds (Fig. 1b, c).

In premotor brain regions, neural activity that supports long-range complex transitions will reflect context information as redundant representations of ongoing behaviour[5–8]. Such representations, referred to here as 'hidden neural states', have been predicted in models of memory-guided behaviour control[9], but are challenging to observe during unconstrained motion in mammals[10–17] or in songbirds with simple syntax rules[18].

Like motor control in many vertebrate species, canary song is governed by a cortico-thalamic loop[19–21] that includes the premotor nucleus HVC[2–4]. In stereotyped songs of zebra finches, HVC projection neurons (PNs) produce stereotyped bursts of activity that are time-locked to song[3]. These cells drive motor outputs or relay timing references to the basal ganglia[22]. In the more variable syllable sequences of Bengalese finches, some PNs fire in a way that depends on neighbouring syllables[18], supporting sequence generation models that include hidden states[9]. However, the time-frame of the song-sequence neural correlations are relatively short (roughly 100 ms). By contrast, correlations in human behaviour can extend for tens of seconds and beyond, and are consistent with long-range syntax rules. At present it is not known whether redundant premotor representations in songbirds can support working memory for syntax control over timescales longer than 100 ms.

To further dissect the mechanisms of working memory for song we used custom head-mounted miniature microscopes to record HVC PNs during song production in freely moving canaries (*Serinus canaria*) (Fig. 2b). Although PNs can be divided into distinct projection-target-specific subtypes, the imaging method does not distinguish these populations and we report results for this mixed population as a whole. These experiments reveal a previously undescribed pattern of neural dynamics that can support structured, context-dependent song transitions and validate predictions of long-range syntax generated by hidden neural states[9,23] in a complex vocal learner.

Complex transitions in a subset of phrases

Inspired by technological advances in human speech recognition[24], we developed a song segmentation and annotation algorithm that automated working with large data sets (more than 5,000 songs; Extended Data Fig. 1a, Methods). The birds' repertoire included 24–37 different syllables with typical durations of 10–350 ms. The average number of syllable repeats per phrase type ranged from 1 to 38, with extreme cases of individual phrases exceeding 10 s and 120 syllables (Extended Data

[1]Department of Biology, Boston University, Boston, MA, USA. [2]Boston University Center for Systems Neuroscience, Boston, MA, USA. [3]Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, CA, USA. [4]Center for Regenerative Medicine of Boston University and Boston Medical Center, Boston, MA, USA. [5]The Pulmonary Center, Boston University School of Medicine, Boston, MA, USA. [6]Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [7]Phil and Penny Knight Campus for Accelerating Scientific Impact, University of Oregon, Eugene, OR, USA. ✉e-mail: ycohen1@mgh.harvard.edu; timg@uoregon.edu

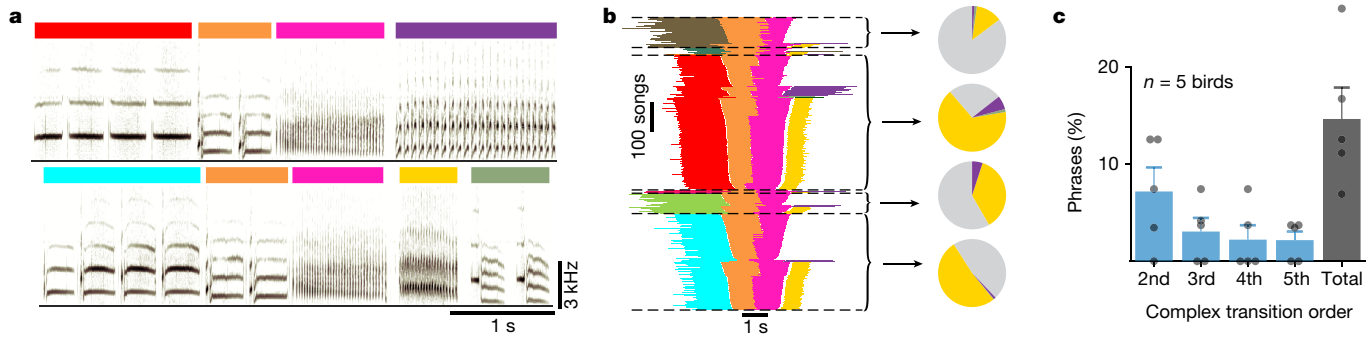


Fig. 1 | Long-range syntax rules in canary song. a, Two example spectrograms of canary song. Coloured bars indicate different phrases assembled from basic elements called syllables. Both examples contain a common phrase transition (orange to pink) but differ in the preceding and following phrases. **b**, A summary of all phrase sequences containing this common transition reveals that the choice of what to sing after the pink phrase depends on the phrases that were produced earlier. Lines represent phrase identity and duration. Song sequences are stacked (vertical axis) and ordered by the identity of the first phrase, the identity of the last phrase, and then the duration of the centre phrases. Pie charts show the frequency of phrases that follow the pink phrase, calculated in the subset of songs that share a preceding sequence context (separated by dashed lines); grey represents the song end and other colours represent a phrase pictured in the left panel. The pink phrase precedes a third-order 'complex transition'; the likelihood that a particular phrase will follow it is dependent on transitions three phrases in the past. **c**, Percentage (mean + s.e.m.) of phrases that precede complex transitions of different orders in $n = 5$ birds (dots).

Fig. 1c–g). Transitions between phrases could be completely deterministic, where one phrase type always followed another, or flexible, where multiple phrase types could follow a given phrase (Fig. 1a, b). In very rare cases, transitions contained an aberrant syllable that could not be stably classified (Extended Data Fig. 2g–i), and all data were visually proofed. (Extended Data Figures 1b and 2 illustrate the reliable annotation of phrase sequences and syllable repertoires.)

As shown in another strain of canaries[1], we found that a small subset of phrase types precede 'complex' transitions—behavioural transitions that depend on the multi-step context of preceding phrases. Specifically, the probability of transition outcomes can change by almost an order of magnitude depending on the identity of the three preceding phrases (Fig. 1b). Such song context dependence is captured by a third-order Markov chain. Extended Data Figure 1i shows the long-range context-dependent transitions for two birds.

HVC neurons encode long-range syntax

To characterize the neural activity that supports complex transitions, we imaged neurons that expressed the genetically encoded calcium indicator GCaMP6f in freely behaving adult male canaries ($n = 3$, age at least one year, recording in left hemisphere HVC[2]). The indicator is selectively expressed in PNs and neural activity to be recorded via fluorescence dynamics extracted from annotated regions of interest (ROIs; Extended Data Fig. 4, Methods). In our data set, 95% of all phrases are trills of multiple syllables and only 6.1% of those are shorter than the decay time constant of the calcium indicator[25] (400 ms; Extended Data Fig. 1h). As in finches, HVC PN activity in canaries was sparse in time[3,18]. Out of $n = 2{,}010$ daily annotated ROIs (mean ± s.d. of 35 ± 15 ROIs per animal per day), about 90% were selectively active in just one or two phrase types (Fig. 2a, c, Extended Data Fig. 5a). This, combined with the long phrase duration (Extended Data Fig. 1f, h), allowed us to examine the song-context dependence of neural activity using GCaMP6f. In our analysis, we treat recordings from different days separately. This approach overestimates the number of independent neurons we imaged but avoids analysis biases and stability concerns. Under the more conservative assumption that sources persist across days, in Supplementary Note 1 we still estimate 1,057 independent sources in our data set.

When we examined the patterns of phrase-locked activity, we identified signals that changed depending on song context. For example, some ROIs showed weak or no activity in one song context but demonstrated strong activity in another song context (Fig. 2a). Notably, this context-dependent activity was strongly influenced by the identity of non-adjacent phrases. For example, Fig. 2d shows the denoised fluorescence signal raster from a ROI, locked to the phrase type marked in pink, which displays a marked variation in activity ($(\Delta f/f_0)_{\text{denoised}}$, Methods) depending on the second phrase in the sequence's past—a second-order correlation. This sequence preference was quantified by integrating the ROI-averaged signal (Extended Data Fig. 5b, c; one-way ANOVA, $F_{5,35} = 18.3$, $P < 1 \times 10^{-8}$; one-way ANOVA evaluates the null hypothesis that there is no activity variation with phrase identity for all sequence-correlated ROIs in this manuscript). We found ROIs with signals that related to the identities of past and future non-adjacent phrases in all three birds (Extended Data Fig. 5). Across all birds, 21.2% of the daily annotated ROIs showed sequence correlations that extended beyond the current active syllable. In 18.1% there were first-order correlations, where activity during one phrase depends on the identity of an adjacent phrase, and in 5.6% there were second-order or greater relations (Extended Data Fig. 5d).

These sequence dependencies could potentially be explained by other factors inherent to the song that may be more predictive of phrase sequence than HVC activity. For example, transition probabilities following a given phrase could potentially depend on the phrase duration[1], on the onset and offset timing of previous phrases, and on the global time since the start of the song—implicating processes such as neuromodulator tone, temperature buildup, or slow adaptation to auditory feedback[26–31] (Extended Data Fig. 6a–g). To rule out these explanations, we used multivariate linear regression and repeated the tests for sequence-correlated neural activity after discounting the effects of these duration and timing variables on the neural signals. We found that 32.8% (39/119 from 3 birds) of second-order or greater relations and 52.7% (147/279 from 3 birds) of first-order relations remained significant (Extended Data Figs. 5c, 6h).

The sequence-correlated ROIs tend to reflect past events more often than future events. Out of $n = 398$ significant correlations between neural activity and phrase sequence, 62.3% reflected preceding phrase identities (binomial z-test rejects the hypothesis of 50%, $z = 6.94$, $P < 1 \times 10^{-11}$). This bias was also found separately in first- or higher-order correlations (Fig. 2e, 60.2% and 67.2%, respectively; both percentages are significantly larger than 50%; binomial z-test, $z = 4.82$, 5.31 and $P < 1 \times 10^{-6}$, $P < 1 \times 10^{-6}$, respectively, and oppose the bias of 44.6% and 43.1% first- and second-order correlations expected to reflect past events from behaviour statistics alone; $P < 1 \times 10^{-7}$, binomial tests) and persisted

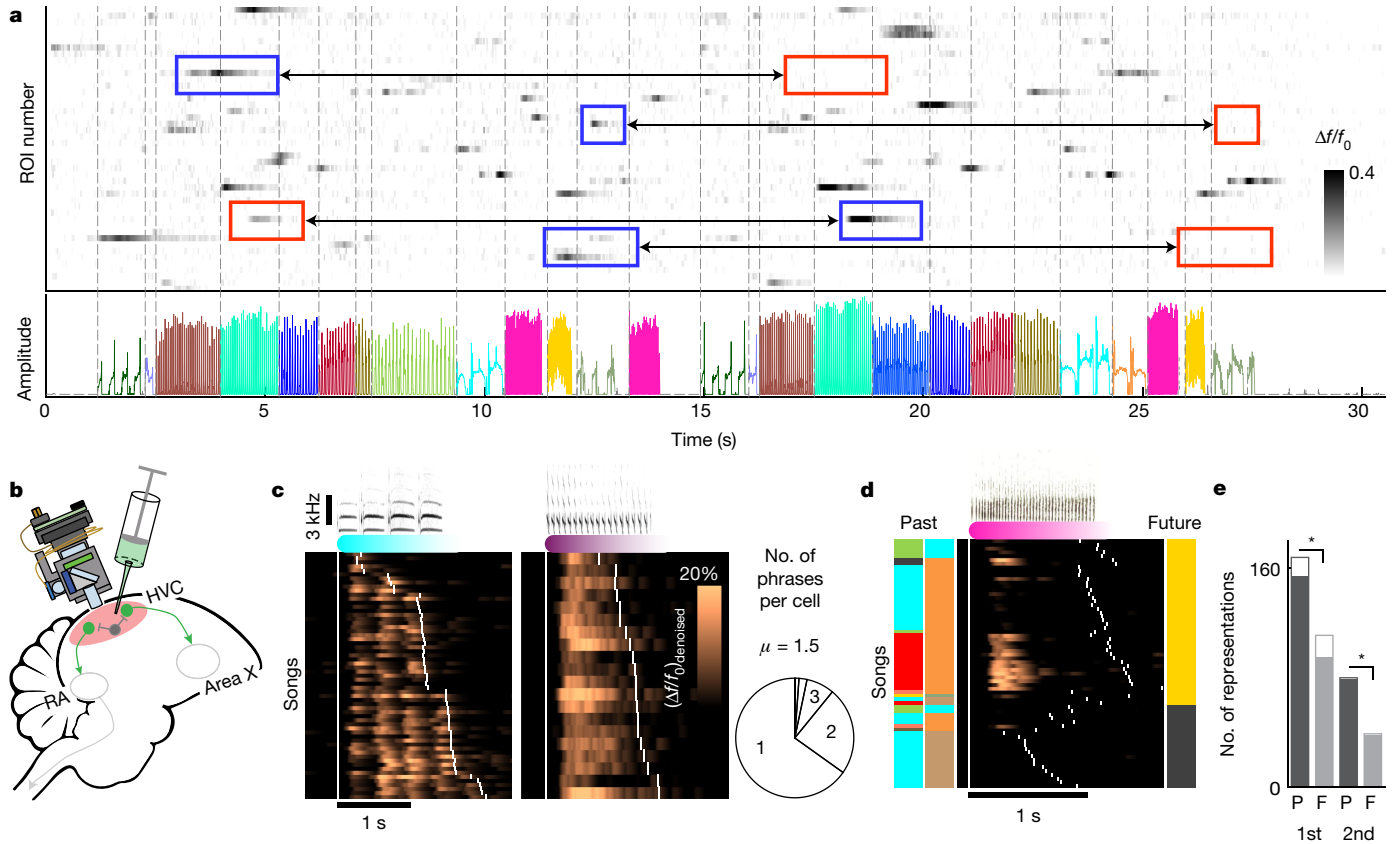


Fig. 2 | HVC PN activity reflects long-range phrase sequence information. **a**, Fluorescence ($\Delta f/f_0$) of multiple ROIs during a singing bout reveals sparse, phrase-type-specific activity. Phrase types are colour coded in the audio amplitude trace (bottom), and dashed lines mark phrase onsets. Context-dependent ROIs show larger phrase-specific signal in one context (blue frames) than another (connected red frames). **b**, Experimental setup. Miniature microscopes were used to image GCaMP6f-expressing neurons in HVC, transduced via lentivirus injection. **c**, Most ROIs are phrase-type-specific. Neural activity is aligned to the onset of phrases. These phrases have long (left) and short (right) syllables and traces are sorted (*y* axis) by phrase duration. White ticks indicate phrase onsets. Pie chart shows fractions of ROIs that are active during just one, two or three phrase types (see Methods). **d**, Phrase-type-specific ROI activity that is strongly related to second upstream phrase identity. Neural activity is aligned to the onset of the current phrase. Songs are arranged by the ending phrase identity (right, colour patches), then by the phrase sequence context (left, colour patches), and then by duration of the pink phrase. White ticks indicate phrase onsets. **e**, Cells reveal more information about past events than future events. Three-hundred and seven different ROIs had 398 significant correlations with adjacent (first order, two left bars) and non-adjacent (second or greater order, two right bars) phrases. The correlations are separated by phrases that precede (P) or follow (F) the phrase, during which the signal is integrated. Empty bars mark transition-locked representations (see Methods, Extended Data Fig. 7d). Two-sided binomial *z*-test to evaluate significant differences (*proportion differences 0.2 ± 0.08 and 0.34 ± 0.11, $z = 4.82$ and 5.31, $P = 1.39 \times 10^{-6}$ and 1.065×10^{-7} for first and second or greater order, respectively).

when we considered ROIs that overlapped in footprint and sequence correlation across days as the same source (Supplementary Note 1). Apart from being more numerous, past correlations also tend to be stronger than future correlations (Extended Data Fig. 6i; significantly larger mean fraction explained variance (η^2) in past correlation, bootstrap comparison rejects the null hypothesis of equal means, $P < 1 \times 10^{-6}$ and $P = 0.001$ for first- and higher-order correlations, respectively).

These findings suggest that, for a subset of HVC neurons, calcium signals are not only related to present motor actions, but also convey the context of past events across multiple syllables.

HVC PNs also encode within-phrase timing

HVC PNs have been recorded in Bengalese finches and in swamp sparrows, two species that also sing strings of syllable repeats. In swamp sparrows, examples of basal ganglia-projecting HVC neurons exhibited stereotyped syllable-locked firing for each syllable in a repeated sequence[32]. In Bengalese finches, the same pattern was described for some cells as well as ramping syllable-locked spike bursts that increased or decreased in spike number over the course of a phrase[18]. In our data set, a small subset of ROIs was consistent with fixed syllable-locked neural activity (Fig. 2c, Extended Data Fig. 7a, c). More commonly, the activity was restricted to a brief period of time within a phrase, as in Fig. 2d, not time-locked to each syllable within the phrase. When we examined all sequence-correlated ROIs, we found that 91% were active for time-intervals shorter than the phrase, with peak timing and onset timing that can be found at all times in the phrase (Fig. 3, Extended Data Fig. 7b, c, e; also showing that some transients could be explained by ramping syllable-locked spike bursts). Together, these findings indicate that the majority of neurons recorded here contain information about timing within a phrase, not just syllable identity.

PNs carry long-range information

Long-range syntax rules imply that a memory of previous elements sung influences future syllable choice. The HVC activity described here provides a clue for a possible mechanism of this process. For example, during a fixed sequence of four phrases, we found ROIs that carried forward information about the identity of the first phrase during each subsequent phrase (Fig. 4a, b, Extended Data Fig. 8a; one-way ANOVA showing significant modulation of neural activity with the identity of the past phrase). In this example, the ROIs that reflect long-range

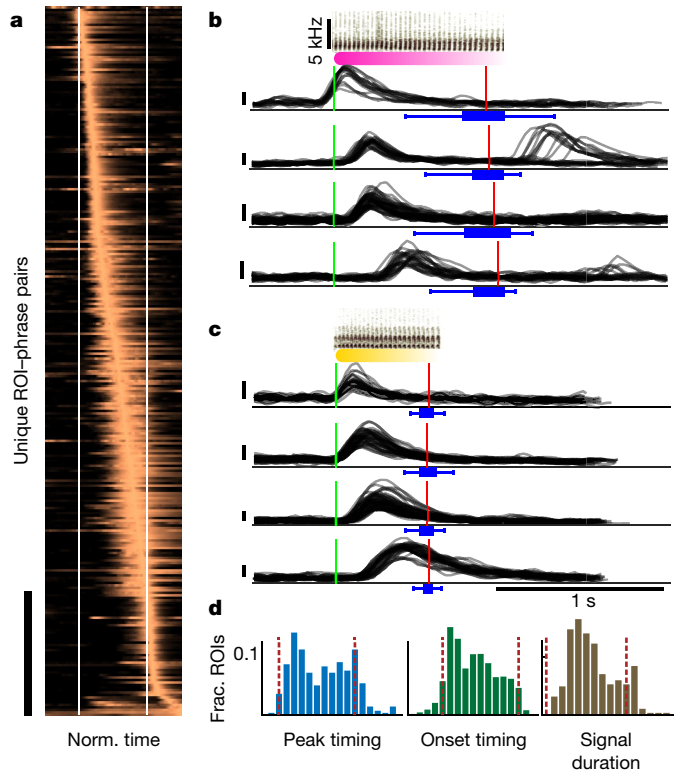


Fig. 3 | Sequence-correlated HVC neurons reflect within-phrase timing. **a**, Activity of context-sensitive ROIs (*y* axis, bar marks 50 rows) is time-warped to fixed phrase edges (*x* axis, white lines) and averaged across repetitions of short-syllable phrases. Traces are ordered by their peak timing to reveal the span of the phrase time frame. **b**, **c**, Example raw $\Delta f/f_0$ traces (*y* axis, vertical bars equal 0.1) of eight ROIs during phrase types that precede (**b**) or follow (**c**) the complex transition in Fig. 1. Traces are aligned to phrase onsets (green line; sonograms show syllables) and panels show ROIs with various onset timing across the phrase. Red lines and blue box plots show the median, range, and quartiles of the phrase offset timing (top to bottom: $n = 70, 23, 55, 39, 40, 38, 50$ and 31 phrases summarized by the box plots). **d**, Histograms showing the distribution of peak timing (left), onset timing (middle) and signal durations (right) of the activity in **a** relative to the phrase edges (dashed lines).

information continue to do so even if the final phrase in the sequence is replaced by the end of the song, suggesting that their activity reflects prior song context rather than some upcoming future syllable choice (Extended Data Fig. 8b; one-way ANOVA, $F_{5,10} = 36.14$ and 2.79, $P < 5 \times 10^{-6}$ and $P < 0.08$ for ROIs 50 and 36, respectively, when replacing the last phrase with the end of song). This example suggests that a chain of neurons that reflect hidden states or information about past choices could provide the necessary working memory to implement long-range transition rules.

HVC neurons active in complex transitions

The phrases in Fig. 4 are phrase types that lead to complex transitions or directly follow them (in Fig. 1). If HVC neurons with context-selective activity are driving long-range syntax rules, then they should represent song context information predominately around complex behaviour transitions, when such information is needed to bias transition probabilities. Accordingly, at the population level, we found more sequence-correlated ROIs around complex transitions; about 70% of sequence-correlated ROIs were found during the rare phrase types that immediately preceded or followed complex transitions (Fig. 4c; 76% (65%) for first (second or greater) order). Both percentages are larger than the 27% (22%) expected from uniform distribution of sequence-correlations in all phrases (binomial test, $P < 1 \times 10^{-10}$, Extended Data Fig. 8c–f) and persist if we consider ROIs that overlap in footprint and sequence correlation across days as the same source (Supplementary Note 1). When we separated the influence of past context and future action on the neural activity we found that, in complex transitions, ROIs predominately represented the identity of the preceding phrase (Extended Data Fig. 8g, h; multi-way ANOVA and Tukey's post hoc analysis showing that the preceding phrase identity significantly affects the neural activity more than twice more often than the following phrase identity; binomial *z*-test rejects the null hypothesis of equal groups: $Z = 6.45$, $P < 1 \times 10^{-10}$). This bias does not occur outside complex transitions (Extended Data Fig. 8i; binomial *z*-test, $Z = 1.06$, $P > 0.1$). This finding suggests that neural coding for past context is enriched during transitions that require this context information.

Ensemble activity predicts complex behaviour

Of the ROIs with first-order and second- or greater-order sequence correlations, 19% and 14%, respectively, were active in several preceding phrase contexts, whereas 44% and 48% preferred just one out of several past contexts (Extended Data Fig. 9). Neurons that respond in multiple contexts can complement each other to provide additional information about song history (Figs. 2d, 4 (ROIs 21, 45, 50)). Extended Data Figure 10a shows four ROIs that were jointly active during a single phrase type. One ROI was active in a single context (ROI 10) and the other three were active in multiple contexts. The phrase during which these ROIs were recorded precedes a complex transition and, in this example, the behaviour alone (prior phrase type) poorly predicts the transition outcome (right bar in Extended Data Fig. 10b, 0.08 out of 1, bootstrapped normalized mutual information estimate; see Methods). However, looking at multiple ROIs together, we found that the network holds significantly more information about the past and future phrase types (Extended Data Fig. 10b, 0.42, 0.33, bootstrapped *z*-test rejects the null hypothesis of equal means, $z = 8.95$, $P < 1 \times 10^{-15}$). This increase exceeds the most informative individual ROIs (0.33, 0.21, bootstrapped *z*-test rejects the null hypothesis of equal means $z = 2.26$, $P < 0.015$ and $z = 5.7$, $P < 1 \times 10^{-8}$, respectively), suggesting synergy of the complementing activity patterns. Furthermore, in this example the network holds more information about the past than the future (Extended Data Fig. 10b–d, bootstrapped *z*-test, $z = 4.32$, $P < 1 \times 10^{-5}$), suggesting that information is lost during the complex transition.

Together, these findings demonstrate that neural activity in canary HVC carries long-range song context information. These hidden states relate primarily (Extended Data Fig. 3) to past or future song and contain the information that is needed to drive complex, context-dependent phrase transitions.

Discussion

Motor skills with long-range sequence dependencies are common in complex behaviours, with speech the richest example. In general, the neural mechanisms that underlie long-range motor sequence dependencies are unknown. Here we show that context-sensitive activity in HVC PNs can support the long-range order in canary song sequences[1]. Specifically, we find PNs the activity of which is contingent on phrases up to four steps in the past and PNs that predict phrases two steps into the future. Cells with this higher-order behaviour tend to be active during complex behavioural transitions—times at which the song behaviour requires high-level information about the sequence context. A key next step will be to further subdivide the activity reported here, in order to determine which PN classes in HVC carry the long-range information.

The HVC activity described here resembles the many-to-one relation between neural activity and behaviour states[9,23,27,33] proposed in some models to relay information across time. In this respect, our findings expand on a previous study in Bengalese finches[18] that identified HVC

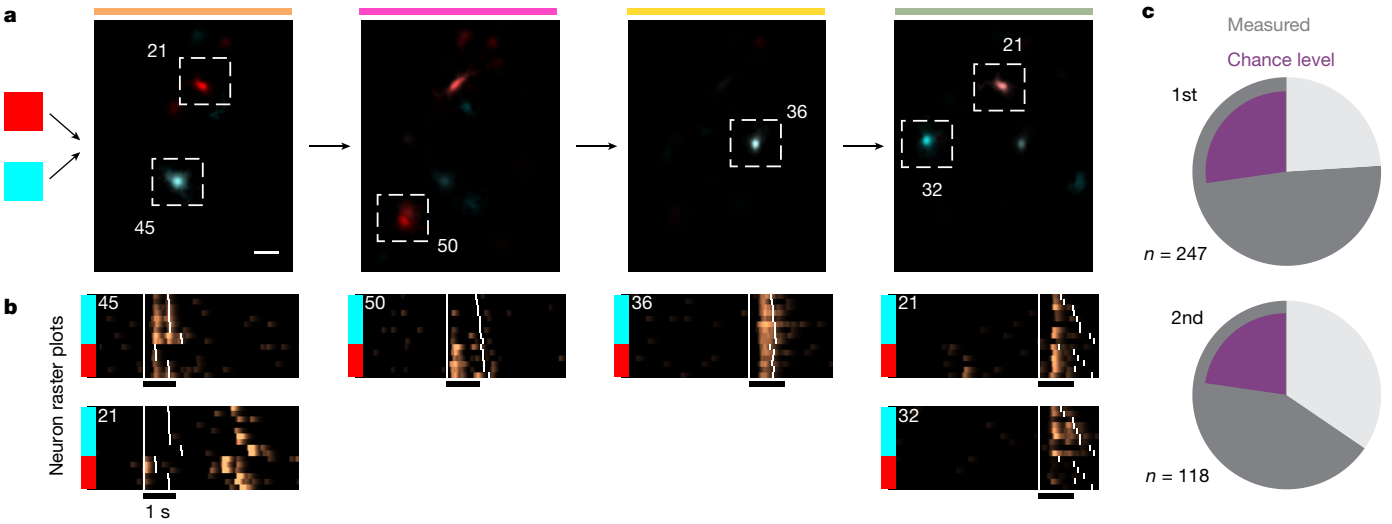


Fig. 4 | Sequence-correlated HVC neurons reflect preceding context up to four phrases apart and show enhanced activity during context-dependent transitions. a, A sequence of four phrases (left to right, colour coded) is preceded by two upstream phrase types (red or cyan). Average maximum projection denoised images (see Methods) are calculated in each sequence context during each phrase in the sequence and overlaid in complementary colours (red, cyan) to reveal context-preferring neurons. Scale bar, 50 µm. **b**, Raster plots of $(\Delta f/f_0)_{\text{denoised}}$ for the ROIs in **a**. Songs are ordered by the preceding phrase type (coloured bars). Extended Data Figure 8a shows the statistical significance of song context relations. Scale bars, 1 s. **c**, Fraction of sequence-correlated ROIs found in complex transitions. Pie charts separate first-order and higher-order sequence correlations. Dark grey summarizes the total fraction for two birds. Purple shows fractions expected from sequence correlates uniformly distributed in all phrase types.

PNs the activity of which depended not just on the current syllable type but also the prior syllable type. This history extended just to the most recent syllable transition, over a time frame of roughly 100 ms.

In the canary HVC neurons observed here, the time frame extends over multiple phrases and several seconds. This longer time frame rules out explanations based on short-term biophysical processes such as short-term calcium dynamics, synaptic plasticity[34], channel dynamics[35] supporting auditory integration[36], sensory–motor delay, and adaptation to auditory inputs[27] that could span a smaller 50–250 ms time frame. Unlike the syllable-locked neural activity reported in Bengalese finches[18], the onset of hidden state activity in canaries is not restricted to phrase edges. Rather, the activity recorded here suggests that parallel chains of sparse neural activity propagate in the song system during a given phrase and that distinct populations of neurons can sequentially encode the same syllable type—a many-to-one mapping of neural sequences onto syllable types that was predicted by a prominent statistical model of birdsong[9].

There are clues that HVC does not contain all of the information required to select a phrase transition—as more neurons correlate to the sequence's past than to its future, it is possible that sequence information in HVC is lost, perhaps owing to neuronal noise that adds stochasticity to transitions. The source of residual stochasticity in HVC could be intrinsic to the dynamics of HVC, resembling the 'noise' terms that are commonly added in sequence generating models[37–39], or may enter downstream, as well-documented noise in the basal ganglia outputs[40] also converges on pre-motor cortical areas downstream of HVC and may affect phrase transitions.

The study of neural dynamics during flexible transitions in canaries may provide a tractable model for studying stochastic cognitive functions—mechanisms in working memory and sensory–motor integration that remain extremely challenging to quantify in most spontaneous behaviours in mammals. Finally, we note that recent marked progress in speech recognition algorithms has used recurrent neural networks with several architectures designed to capture sequence dependencies with hidden states. Examples include long short-term memory (LSTM)[41], hierarchical time scales[42], hidden memory relations[43], and attention networks[44]. It is possible that machine learning models will help to interpret the complex dynamics of the song system and to inform new models of many-to-one, history-dependent mappings between brain state and behaviour[23].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2397-3.

1. Markowitz, J. E., Ivie, E., Kligler, L. & Gardner, T. J. Long-range order in canary song. *PLOS Comput. Biol.* **9**, e1003052 (2013).
2. Nottebohm, F., Stokes, T. M. & Leonard, C. M. Central control of song in the canary, *Serinus canarius*. *J. Comp. Neurol.* **165**, 457–486 (1976).
3. Hahnloser, R. H. R., Kozhevnikov, A. A. & Fee, M. S. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* **419**, 65–70 (2002).
4. Long, M. A. & Fee, M. S. Using temperature to analyse temporal dynamics in the songbird motor pathway. *Nature* **456**, 189–194 (2008).
5. Rokni, U., Richardson, A. G., Bizzi, E. & Seung, H. S. Motor learning with unstable neural representations. *Neuron* **54**, 653–666 (2007).
6. Todorov, E. Optimality principles in sensorimotor control. *Nat. Neurosci.* **7**, 907–915 (2004).
7. Wolpert, D. M. Computational approaches to motor control. *Trends Cogn. Sci.* **1**, 209–216 (1997).
8. Leonardo, A. Degenerate coding in neural systems. *J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol.* **191**, 995–1010 (2005).
9. Jin, D. Z. & Kozhevnikov, A. A. A compact statistical model of the song syntax in Bengalese finch. *PLOS Comput. Biol.* **7**, e1001108 (2011).
10. Ohbayashi, M., Ohki, K. & Miyashita, Y. Conversion of working memory to motor sequence in the monkey premotor cortex. *Science* **301**, 233–236 (2003).
11. Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14**, 477–485 (1995).
12. Svoboda, K. & Li, N. Neural mechanisms of movement planning: motor cortex and beyond. *Curr. Opin. Neurobiol.* **49**, 33–41 (2018).
13. Thompson, J. A., Costabile, J. D. & Felsen, G. Mesencephalic representations of recent experience influence decision making. *eLife* **5**, e16572 (2016).
14. Pastalkova, E., Itskov, V., Amarasingham, A. & Buzsáki, G. Internally generated cell assembly sequences in the rat hippocampus. *Science* **321**, 1322–1327 (2008).
15. Churchland, M. M., Afshar, A. & Shenoy, K. V. A central source of movement variability. *Neuron* **52**, 1085–1096 (2006).
16. Mushiake, H., Saito, N., Sakamoto, K., Itoyama, Y. & Tanji, J. Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. *Neuron* **50**, 631–641 (2006).

17. Shima, K. & Tanji, J. Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *J. Neurophysiol.* **84**, 2148–2160 (2000).
18. Fujimoto, H., Hasegawa, T. & Watanabe, D. Neural coding of syntactic structure in learned vocalizations in the songbird. *J. Neurosci.* **31**, 10023–10033 (2011).
19. Hamaguchi, K., Tanaka, M. & Mooney, R. A distributed recurrent network contributes to temporally precise vocalizations. *Neuron* **91**, 680–693 (2016).
20. Ashmore, R. C., Wild, J. M. & Schmidt, M. F. Brainstem and forebrain contributions to the generation of learned motor behaviors for song. *J. Neurosci.* **25**, 8543–8554 (2005).
21. Alonso, R. G., Trevisan, M. A., Amador, A., Goller, F. & Mindlin, G. B. A circular model for song motor control in *Serinus canaria*. *Front. Comput. Neurosci.* **9**, 41 (2015).
22. Goldberg, J. H. & Fee, M. S. Singing-related neural activity distinguishes four classes of putative striatal neurons in the songbird basal ganglia. *J. Neurophysiol.* **103**, 2002–2014 (2010).
23. Jin, D. Z. Generating variable birdsong syllable sequences with branching chain networks in avian premotor nucleus HVC. *Phys. Rev. E* **80**, 051902 (2009).
24. Hinton, G. et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* **29**, 82–97 (2012).
25. Chen, T.-W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
26. Bouchard, K. E. & Brainard, M. S. Auditory-induced neural dynamics in sensory–motor circuitry predict learned temporal and sequential statistics of birdsong. *Proc. Natl Acad. Sci. USA* **113**, 9641–9646 (2016).
27. Wittenbach, J. D., Bouchard, K. E., Brainard, M. S. & Jin, D. Z. An adapting auditory-motor feedback loop can contribute to generating vocal repetition. *PLOS Comput. Biol.* **11**, e1004471 (2015).
28. Dave, A. S., Yu, A. C. & Margoliash, D. Behavioral state modulation of auditory activity in a vocal motor system. *Science* **282**, 2250–2254 (1998).
29. Cardin, J. A. & Schmidt, M. F. Noradrenergic inputs mediate state dependence of auditory responses in the avian song system. *J. Neurosci.* **24**, 7745–7753 (2004).
30. Glaze, C. M. & Troyer, T. W. Development of temporal structure in zebra finch song. *J. Neurophysiol.* **109**, 1025–1035 (2013).
31. Castelino, C. B. & Schmidt, M. F. What birdsong can teach us about the central noradrenergic system. *J. Chem. Neuroanat.* **39**, 96–111 (2010).
32. Prather, J. F., Peters, S., Nowicki, S. & Mooney, R. Precise auditory–vocal mirroring in neurons for learned vocal communication. *Nature* **451**, 305–310 (2008).
33. Okubo, T. S., Mackevicius, E. L., Payne, H. L., Lynch, G. F. & Fee, M. S. Growth and splitting of neural sequences in songbird vocal development. *Nature* **528**, 352–357 (2015).
34. Zucker, R. S. & Regehr, W. G. Short-term synaptic plasticity. *Annu. Rev. Physiol.* **64**, 355–405 (2002).
35. Iacobucci, G. J. & Popescu, G. K. NMDA receptors: linking physiological output to biophysical operation. *Nat. Rev. Neurosci.* **18**, 236–249 (2017).
36. Nagel, K., Kim, G., McLendon, H. & Doupe, A. A bird brain's view of auditory processing and perception. *Hear. Res.* **273**, 123–133 (2011).
37. Fiete, I. R., Senn, W., Wang, C. Z. H. & Hahnloser, R. H. R. Spike-time-dependent plasticity and heterosynaptic competition organize networks to produce long scale-free sequences of neural activity. *Neuron* **65**, 563–576 (2010).
38. Abeles, M. *Corticonics: Neural Circuits of the Cerebral Cortex* (Cambridge Univ. Press, 1991).
39. Cannon, J., Kopell, N., Gardner, T. & Markowitz, J. Neural sequence generation using spatiotemporal patterns of inhibition. *PLOS Comput. Biol.* **11**, e1004581 (2015).
40. Hamaguchi, K. & Mooney, R. Recurrent interactions between the input and output of a songbird cortico-basal ganglia pathway are implicated in vocal sequence variability. *J. Neurosci.* **32**, 11671–11687 (2012).
41. Graves, A., Mohamed, A. & Hinton, G. Speech recognition with deep recurrent neural networks. *2013 IEEE Intl Conf. Acoustics, Speech and Signal Processing* 6645–6649 (2013).
42. Yamashita, Y. & Tani, J. Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLOS Comput. Biol.* **4**, e1000220 (2008).
43. Santoro, A. et al. in *Advances in Neural Information Processing Systems 31* (eds Bengio, S. et al.) 7310–7321 (Curran Associates, 2018).
44. Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K. & Bengio, Y. in *Advances in Neural Information Processing Systems 28* (eds Cortes, C. et al.) 577–585 (Curran Associates, 2015).

Methods

Ethics declaration

All procedures were approved by the Institutional Animal Care and Use Committee of Boston University (protocol numbers 14-028 and 14-029).

Birds

Imaging data were collected from $n = 3$ adult male canaries. Birds were individually housed for the entire duration of the experiment and kept on a light–dark cycle matching the daylight cycle in Boston (42.3601° N) with unlimited access to food and water. The sample sizes in this study are similar to sample sizes used in the field. The birds were not used in any other experiments. This study did not include experimental groups and did not require blinding or randomization.

Surgical procedures

Anaesthesia and analgesia. Before the birds were anaesthetized, they were injected with meloxicam (intramuscular, 0.5 mg/kg) and deprived of food and water for a minimum of 30 min. Birds were anaesthetized with 4% isoflurane and maintained at 1–2% for the course of the surgery. Prior to skin incision, bupivacaine (4 mg/kg in sterile saline) was injected subcutaneously (volume 0.1–0.2 ml). Meloxicam was also administered for 3 days after surgery.

Stereotactic coordinates. The head was held in a previously described, small animal stereotactic instrument[45]. To increase anatomical accuracy and ease of access, we deviated from the published atlas coordinates[45] and adapted the head angle reference to a commonly used forehead landmark parallel to the horizontal plane. The outer bone leaflet above the prominent λ sinus was removed and the medial (positive = right) and anterior (positive) coordinates are measured from that point. The depth is measured from the brain's dura surface. The following coordinates were used (multiple values indicate multiple injections): HVC: +65°, −2.5 mm ML, 0.12 mm AP, 0.15–0.7 mm D; nucleus RA: +80°, −2.5 mm ML, −1.2 mm AP, 1.9–3 mm D; area X: +20°, −1.27, −1.3 mm ML, 5.65, 5.8 mm AP, 2.65–2.95 mm D. Angles are measured from the horizontal plane defined above and increase as the head is rotated downward, the mediolateral coordinate (ML) is measured from the midline and increases rightward, the anterior–posterior coordinate (AP) is parallel to the horizontal plane and measured forward from λ, and the depth (D) is measured from the brain's surface and increases with depth.

HVC demarcation and head anchoring. To target HVC, 50–100 nl of the retrograde lipophilic tracer DiI (5 mg/ml solution in dimethylformamide, DMF) was injected into the left area X. The outer bone leaflet was removed above area X using a dental drill. The inner bone leaflet was thinned and removed using an ophthalmic scalpel, exposing a hole of ~300 μm diameter. The left area X was injected using a Drummond Nanoject II (Drummond pipette, 23 nl/s, pulses of 2.3 nl). In the same surgery, a head anchoring structure was created by curing dental acrylic (Flow-It ALC, Pentron) above the exposed skull and through ~100-μm holes in the outer bone leaflet.

Virus injection and lens implants. A lentivirus that was developed for previous work in zebra finches (containing the vector pHAGE-RSV-GCaMP6f; Addgene plasmid 80315) was also used in canaries[46]. The outer skull leaflet above HVC was removed with a dental drill. The inner bone leaflet was thinned and removed with an ophthalmic scalpel, exposing an area of the dura about 1.5–2 mm in diameter. The DiI demarcation of HVC was used to select an area for imaging. The lentivirus was injected in 3 or 4 locations, at least 0.2 mm apart, at a range of depths between 0.5 and 0.15 mm. In total 800–1,000 nl was injected into the left HVC. After the injection, the dura was removed and the parahippocampus segment above the imaging site was removed using a dura pick and a custom tissue suction nozzle. A relay GRIN lens (Grintech GT-IFRL-100, 0.44 pitch length, 0.47 NA) was immediately positioned on top of the exposed HVC and held in place with Kwik-Sil (WPI). Dental acrylic (Flow-It, Pentron) was used to attach the lens to the head plate and to cover the surgery area. The birds were allowed to recover for 1–2 weeks.

Hardware

To image calcium activity in HVC PNs during singing, we used custom, lightweight (~1.8 g), commutable, 3D-printed, single-photon head-mounted fluorescent microscopes that simultaneously record audio and video (Fig. 2). These microscopes enabled us to record hundreds of songs per day, and all songs were recorded from birds longitudinally in their home cage, without requiring adjustment or removal of the microscope during the imaging period. Birds were imaged for less than 30 min total on each imaging day, and LED activation and video acquisition were triggered on song using previously described methods[46].

Microscope design. We used a custom, open-source microscope developed in the lab[46]. A blue LED produces excitation light (470-nm peak, LUXEON Rebel). A drum lens collects the LED emission, which passes through a 4 mm × 4 mm excitation filter, deflects off a dichroic mirror, and enters the imaging pathway via a 0.25 pitch gradient refractive index (GRIN) objective lens. Fluorescence from the sample returns through the objective, the dichroic, an emission filter, and an achromatic doublet lens that focuses the image onto an analogue CMOS sensor with 640 × 480 pixels mounted on a PCB that also integrates a microphone. The frame rate of the camera is 30 Hz, and the field of view is approximately 800 μm × 600 μm. The housing is made of 3D-printed material (Formlabs, black resin). A total of five electrical wires run out from the camera: one wire each for camera power, ground, audio, NTSC analogue video and LED power. These wires run through a custom flex-PCB interconnect (Rigiflex) up to a custom-built active commutator. The NTSC video signal and analogue audio are digitized through a USB frame-grabber. Custom software written in the Swift programming language running on the macOS operating system (version 10.10) leverages native AVFoundation frameworks to communicate with the USB frame-grabber and capture the synchronized audio–video stream. Video and audio are written to disk in MPEG-4 container files with video encoded at full resolution using either H.264 or lossless MJPEG Open DML codecs and audio encoded using the AAC codec with a 48-kHz sampling rate. All schematics and code can be found online https://github.com/gardner-lab/FinchScope and https://github.com/gardner-lab/video-capture.

Microscope positioning and focusing. Animals were anaesthetized and head fixed. The miniaturized microscope was held using a manipulator and positioned above the relay lens. The objective distance above the relay was set such that blood vessels and GCaMP6f expressing cells were in focus. The birds recovered in the recording setup. Within the first couple of weeks, the microscopes were refocused to maximize the number of observable neurons.

Histological verification of genetic tool properties

DiI was injected into area X as described above. Three days later, ~800 nl lentivirus was injected into HVC using the DiI demarcation. In finches, this virus infected predominately PNs[46]. In this project we analysed neurons with sparse activity that do not match the tonic activity of interneurons in HVC. The virus was injected into four sites, at least 0.2 mm apart and at two depths (matching the in-vivo imaging experiment's procedure above). About four weeks later, the bird was euthanized (by intracoelomic injection of 0.2 ml 10% Euthasol; Virbac, ANADA 200-071, in saline) and perfused by first running saline and then 4% paraformaldehyde via the heart's left chamber and the contralateral neck vein. The brain was extracted and kept overnight in 4% paraformaldehyde at 4 °C.

GCaMP6f expression. The fixed tissue was sectioned into 70-μm sagittal slices (Vibratome series 1000), placed on microscope slides, and sealed with cover slips and nail polish. Epifluorescence images were taken using a Nikon Eclipse Ni-E tabletop microscope (Extended Data Fig. 4a).

Expression specificity to excitatory neurons. The fixed tissue was immersed in 20% sucrose solution overnight and then 30% sucrose solution over the following night, frozen and sectioned into 30-μm sagittal slices (Cryostat, Leica CM3050S). Following work in zebra finches[47], the slices were stained using antibodies against the calcium binding interneuron markers calbindin (1:4,000, SWANT), calretinin (1:15,000, SWANT), and parvalbumin (1:1,000, SWANT) by overnight incubation with the primary antibody at 4 °C and with a secondary antibody (coupled to Alexa Fluor 647) for 2 h at room temperature. Slices were mounted on microscope slides and sealed with cover slips and nail polish. A confocal microscope (Nikon C2si) was used to image GCaMP6f and the interneuron markers in 3-μm-thick sections through the tissue (Extended Data Fig. 4b). The images were inspected for co-stained cells (for example, see Supplementary Videos 1–7). The results ruled out any co-expression of GCaMP and calbindin or calretinin. We found two cells that expressed both parvalbumin and GCaMP (Supplementary Video 5 shows one example; <0.5% of parvalbumin-stained cells, <0.01% of GCaMP-expressing cells), possibly replicating a previous observation of parvalbumin expression in HVC PNs[47].

Data collection

Song screening. Birds were individually housed in soundproof boxes and recorded for 3–5 days (Audio-Technica AT831B Lavalier Condenser Microphone, M-Audio Octane amplifiers, HDSPe RayDAT sound card and VOS Games' Boom Recorder software on a Mac Pro desktop computer). In-house software was used to detect and save only sound segments that contained vocalizations. These recordings were used to select subjects that were copious singers (≥50 songs per day) and produced at least 10 different types of syllable.

Video and audio recording. All data used in this manuscript were acquired between late February and early July—a period during which canaries perform their mating season songs. To avoid overexposure of the fluorescent proteins, data collection was done during the morning hours (from sunrise until about 10 am) and the daily accumulated LED-on time rarely exceeded 30 min. Audio and video data collection was triggered by the onset of song as previously described[46] with an additional threshold on the spectral entropy that improved the detection of song periods markedly. Data files from the first couple of weeks, a period during which the microscope focusing took place and the birds sang very little, were not used. Additionally, data files from (extremely rare) days on which video files were corrupted because of tethering malfunctions were not used.

Data analysis

Video file preprocessing. Software developed in-house was used to load video frames and audio signal to MATLAB (https://github.com/gardner-lab/FinchScope/tree/master/Analysis%20Pipeline/extract-media) along with the accompanying timestamps. Video frames were interpolated in time and aligned to an average frame rate of 30 Hz. Audio samples were aligned and trimmed in sync with the interpolated frame timestamps. To remove out-of-focus bulk fluorescence from the 3D representation of the video (rows × columns × frames), the background was subtracted from each frame by smoothing it with a 145-pixel-wide circular Gaussian kernel, resulting in 3D video data, $V(x,y,t)$.

Audio processing. Song syllables were segmented and annotated by a semi-automatic process. First, a set of ~100 songs was manually annotated using a GUI developed in-house (https://github.com/yardencsGitHub/BirdSongBout/tree/master/helpers/GUI). This set was chosen to include all potential syllable types as well as cage noises. The manually labelled set was then used to train a deep learning algorithm ('TweetyNet') developed in-house (https://github.com/yardencsGitHub/tweetynet). The trained algorithm annotated the rest of the data and its results were manually verified and corrected. In both the training phase of TweetyNet and the prediction phase for new annotations, data were fed to TweetyNet in segments of 1 s and the output of TweetyNet was the most likely label for each 2.7-ms time bin in the recording.

Assuring the separation of syllable classes. To make sure that the syllable classes were well separated, all the spectrograms of every instance of every syllable, as segmented in the previous section, were zero-padded to the same duration, pooled and divided into two equal sets. For each pair of syllable types, a support vector machine classifier was trained on half the data (the training set) and its error rate was calculated on the other half (the test set). These results are presented, for example, in Extended Data Fig. 1b.

Testing for within-class context distinction by syllable acoustics. Apart from the clear between-class separation of different syllables for syllables that precede complex transitions, we checked the within-class distinction between contexts that affect the transition. To do that, we used previously published parameters[48] and treated each syllable rendition as a point in an eight-dimensional space of normalized acoustic features. For a pair of syllable groups (different syllables or the same syllable in different contexts) we calculate the discriminability coefficient:

$$d'_{\mathrm{AB}} = \frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{2} + \frac{\sigma_B^2}{2}}}$$

Where $\mu_A - \mu_B$ is the L_2 distance between the centres of the distributions and σ_A^2 and σ_B^2 are the within-group distance variances from the centres. Extended Data Figure 3 demonstrates that all within-class d' values are smaller than all between-class d' values.

Identifying complex transitions. Complex transitions were identified by the length of the Markov chain required to describe the outcome probabilities. These dependencies were found using a previously described algorithm that extracts the probabilistic suffix tree[1] (PST) for each transition (https://github.com/jmarkow/pst). In brief, the tree is a directed graph in which each phrase type is a root node that represents the first-order (Markov) transition probabilities to downstream phrases, including the end of song. The pie chart in Extended Data Fig. 1i (i) shows such probabilities. Upstream nodes represent higher-order Markov chains (2nd and 3rd in Extended Data Fig. 1i (ii) and (iii), respectively) that are added sequentially if they significantly add information about the transition.

ROI selection, Δf/f signal extraction and de-noising. Song-containing movies were converted to images by calculating, for each pixel, the maximal value across all frames. These 'maximum projection images' were then similarly used to create a daily maximum projection image and also concatenated to create a video. The daily maximum projection and song-wise maximum projection videos were used to select regions of interest (ROIs), purported single neurons, in which fluorescence fluctuated across songs.

ROIs were never smaller than the expected neuron size, did not overlap, and were restricted to connected shapes that rarely deviated from simple ellipses. Notably, this selection method did not differentiate between sources of fixed and fluctuating fluorescence. The footprint of each ROI in the video frames was used to extract the time series, $f(t) = \sum_{(x,y)\in \mathrm{ROI}} V(x,y,t)$, summing signal from all pixels within that

ROI. Then, signals were converted to relative fluorescence changes, $\frac{\Delta f(t)}{f_0} = \frac{f(t)-f_0}{f_0}$ by defining f_0 to be the 0.05 quantile.

The denoised fluorescence, $(\Delta f/f_0)_{\text{denoised}}$, was estimated from the relative fluorescence change using previously published modelling of the calcium concentration dynamics and the added noise process caused by the fluorescence measurement[49].

Seeking ROIs with sequence correlations. As each ROI was sparsely active in very few phrase types, we first sought ROIs that were active during a phrase type and then tested whether it showed correlations to preceding or following phrase identities. We used the following two-step scheme.

Step 1: identify ROIs with phrase-type-active signal. Phrase-type-active ROI was defined by requiring signal, $s(t) = \frac{\Delta f(t)}{f_0}$ as defined in the previous section, to be larger and distinct from noise fluctuations (for each ROI and repeats of each phrase type, P). The 0.9 quantile, Δff_{90}, was taken as a measure of within-phrase peak values to reduce outliers. Irrespective of the phrase boundaries, periods of time during which an ROI was active were separated from baseline noise fluctuations by fitting the signal within an ROI, $s(t)$, with a two-state hidden Markov model with Gaussian emission functions. Specifically, at time t the observable, $s(t)$, is assumed to follow a Gaussian distribution, $\mathbb{N}(\mu_t,\sigma_t)$, that determines the likelihood $p(s(t);\mu_t,\sigma_t)$. The hidden variable, $\mathbf{\Theta}_t = (\mu_t,\sigma_t)$, is defined by the mean ($\boldsymbol{\mu} = \mu_1, \mu_2$) and standard deviation ($\boldsymbol{\sigma} = \sigma_1, \sigma_2$) of the Gaussian distributions and follows first-order time-independent Markov transition probabilities, $R = p(\mathbf{\Theta}_{t+1}|\mathbf{\Theta}_t)$, a 2×2 matrix of transition probabilities between two states ('activity' and 'noise'). To estimate the sequence of states (the hidden process $\mathbf{\Theta}$), we maximize the log-likelihood: $L\{\mathbf{s},\mathbf{\Theta},R,\boldsymbol{\mu},\boldsymbol{\sigma}\} = \sum_t \log p(\mathbf{\Theta}_t|\mathbf{\Theta}_{t-1}) + \sum_t \log p(s(t)|\mathbf{\Theta}_t)$. In this process, the mean ($\boldsymbol{\mu}$) and standard deviation ($\boldsymbol{\sigma}$) of the two Gaussian distributions are free parameters.

We define the phrase-type-occupancy, HMM_P, as the fraction of phrase P repetitions that contained the 'active' state. These two activity measures, Δff_{90} and HMM_P, are used to select ROIs to be investigated for sequence correlations. We impose lenient thresholds: $\Delta ff_{90} > 0.1$ (that is, fluorescence fluctuation is larger than a 10% deviation from baseline); and $\text{HMM}_P > 0.1$ (that is, the phrase type carries neural activity in 10% of occurrences or more). In our data set, this threshold is roughly equivalent to ignoring ROIs that are active only once or twice during a recording day.

Step 2: test sequence correlations. First-order relationships between the signal integral (summed across time bins in the phrase) and the upstream or downstream phrase identities were tested using a one-way ANOVA. The entire set of songs for each bird was used to calculate the first-order phrase transition probabilities, $P_{ab} = P(a \to b)$, for all phrases a and b. Second-order relationships were tested between the signal integral and the identity of the second upstream (downstream) phrase identity for all intermediate phrase types that preceded (followed) the phrase-in-focus in at least 10% of the repeats (as indicated by the phrase transition matrix). Sequence–signal correlations were not investigated if fewer than $n = 10$ repeats contributed to the test. Relations were discarded if the label that led to the significant ANOVA contained only one song. Data used for ANOVA tests are represented in Extended Data figures by box plots marking the median (centre line); upper and lower quartiles (box limits); extreme values (whiskers), and outliers (+ markers).

The data were not tested for normality before performing ANOVA tests for individual neurons with the following reasoning. Statistics textbooks suggest that violating the normality requirement is not expected to have a significant effect. For example, Howell[50] writes: "As we have seen, the analysis of variance is based on the assumptions of normality and homogeneity of variance. In practice, however, the analysis of variance is a robust statistical procedure, and the assumptions frequently can be violated with relatively minor effects. This is especially true for the normality assumption. For studies dealing with this problem, see Box (1953, 1954a, 1954b))), Boneau (1960), Bradley (1964), and Grissom (2000)." In addition, carrying tests for normality will create a bias in our analyses. Each neuron that is tested for phrase sequence correlation is recorded in a different number of songs. Testing for normality will create a bias towards larger numbers of songs and against high-order correlations.

Nevertheless, we repeated the analyses in this manuscript with non-parametric one-way ANOVA (Kruskal–Wallis). Although ~15% fewer neurons passed the more stringent tests, all the results in this article remained the same. We include a summary of the non-parametric statistics as Supplementary Note 2.

Note that, in this procedure, sparsely active ROIs or ROIs that were active in rare phrase types were not tested for sequence correlation. In the main text we reported that 21.2% of the entire set of ROIs showed sequence correlation. This percentage includes ROIs that were not tested for sequence correlations. Out of the ROIs that were tested, about 30% had significant sequence correlations (23% and 10% showed first- and second-order correlations).

Phrase specificity. The fraction of phrase repetitions during which a ROI is 'active', HMM_P, was also used to calculate the phrase specificity of an ROI (Fig. 2). For each ROI, the fraction of activity in repetitions of each phrase was calculated separately. These measures were normalized and sorted in descending order. Then, the number of phrase types that accounted for 90% of the ROI's activity was calculated.

Transition-locked activity onsets. The hidden Markov modelling of neural activity was used to identify signal onsets at transition from the 'noise' to the 'active' states (Fig. 2e, Extended Data Fig. 7d). The phrase transition segment is defined as the time window between the onset of the last syllable in one phrase and the offset of the first syllable in the next phrase. ROIs for which the sequence-correlated activity initiated during the phrase transition in the majority of cases were suspected as transition-locked representations. These activity rasters were manually examined and a small number of representations (nine) were excluded from population-level statistics because they appeared reliably and exclusively in specific transitions. Signals that occur exclusively in specific transitions are trivially sequence correlated but simply reflect the ongoing behaviour. This exclusion does not change the results in this paper.

Controlling for phrase durations and time-in-song confounds. In songs that contain a fixed phrase sequence, as in Fig. 2d, we calculated the significance of the relation between $s = \sum_{t\in P}(\Delta f/f_0)_{\text{denoised}}$, an integral of the signal during one phrase in the sequence, the target phrase P, and the identity of an upstream (or downstream) phrase that changes from song to song using a one-way ANOVA. This relation can be carried by several confounding variables: the duration of the target phrase; the relative timing of intermediate phrase edges, between the changing phrase and the target phrase; and the absolute time-in-song of the target phrase.

In Extended Data Fig. 6h we account for these variables by first calculating the residuals of a multivariate linear regression (a general linear model, or GLM) between those variables and s, and then using a one-way ANOVA to test the relation of the residuals and the upstream or downstream phrase identity.

Comparing numbers of significant sequence correlations to past and future events. In Fig. 2e, we compare the numbers of significant sequence correlations between two groups. Group sizes were converted to fractions and the binomial comparison z-statistic was used to compare those fractions. Generally, the statistic $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ with $\hat{p}_1, \hat{p}_2$ the measured fractions of significant correlations in two

populations of sizes n_1, n_2 and $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ is tested against the normal distribution null hypothesis of zero mean. The effect size, $\hat{p}_1 - \hat{p}_2$, has the confidence interval $\mathrm{CI} = \pm 1.96 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$.

In this comparison there is no bias from the conditions of the statistical test (one-way ANOVA) used to establish sequence correlations of individual ROIs. The process of seeking ROIs with sequence correlations (described above) guarantees that tests were not carried in under-sampled conditions because the minimal number of repetitions always exceeded the number of song contexts. In these conditions the ANOVA test is not biased by the number of song contexts, or branching order, in different transitions because the test's significance threshold depends on the number of statistical degrees of freedom that account for the number of contexts. This dependence guarantees that tests with more (or fewer) song contexts are not more likely to reach statistical significance by chance.

Contrasting the strength of sequence correlation to past and future events. For one-way ANOVA tests, we estimated the significance of the difference in η^2-statistics (fraction explained variance) calculated in past versus future correlations using the following bootstrapping procedure. First, we pooled all η^2-statistics together. Then we randomly split the pool into 'past' and 'future' groups of the same size as the data in Fig. 2e and calculated the mean value in each group. We repeated this process 1,000,000 times and used this bootstrapped distribution to calculate a P value for the original difference between means. This process was carried out separately for first-order sequence correlations and for second-order or greater sequence correlations (Extended Data Fig. 6i).

Peak location, onset location, and relative duration of sequence correlated activity. The data in Fig. 3a were used to create the following three distributions (Fig. 3d). 1, Relative peak timing: the trial-averaged signals (rows in Fig. 3a differ in ROIs and phrase type) were calculated after time-warping the signals to a fixed phrase duration, $T_{\mathrm{phrase}} = 1$, the onset of which is set to $T_{\mathrm{onset}} = 0$. The timing of the signal peak, t_{peak}, is therefore already normalized because $t_{\mathrm{peak}} = (t_{\mathrm{peak}} - T_{\mathrm{onset}})/T_{\mathrm{phrase}}$.

2, Relative onset timing: the signal in each trial that contributed to Fig. 3a was fitted with a hidden Markov model (as explained in 'Seeking ROIs with sequence correlations'). The onset time point of the signal state, t_{onset}, was normalized with respect to the phrase onset time, T_{onset}, and the phrase duration, T_{phrase}:

$$\hat{t}_{\mathrm{onset}} = \frac{t_{\mathrm{onset}} - T_{\mathrm{onset}}}{T_{\mathrm{phrase}}}.$$

3, Relative signal duration: a threshold at 0.5 was used to identify segments of reliable state occupancy within the traces in Extended Data Fig. 7d. The resulting signal segments are in time-normalized coordinates and represent the duration relative to the phrase duration.

Simulating point neuron fluorescence response to spike trains. To simulate the expected calcium indicator signal in response to a spike train, $sp(t)$ (Extended Data Fig. 7a), we used the empirical single-spike response:

$$K(t) = \begin{cases} \frac{1 - \mathrm{e}^{-t/0.045}}{1 - \mathrm{e}^{-1}} & 0 \le t \le 0.045\mathrm{s} \\ \mathrm{e}^{-(t-0.045)/0.142} & t > 0.045\mathrm{s} \end{cases}$$

Corresponding to a rise time constant of 45 ms and a decay time constant of 142 ms (see supplementary table 3 in ref. [25]). The above kernel is a low boundary on the rise time because it assumes 45 ms for the full signal rise time and not just half-way. This is done to give a limit on what can be resolved.

For a point neuron, we do not assume other dynamical processes that stem from morphology. The simulated signal is the convolution of the spike train with the kernel, K:

$$F(t) = \int_{-\infty}^{t} sp(\tau)K(t-\tau)\mathrm{d}\tau$$

Contrasting influence of preceding and following phrases on neural activity. For neurons with significant sequence correlations (one-way ANOVA, described above), we adopted a method agnostic to correlation order (first or higher, as defined above) and direction (past or future) (Extended Data Fig. 8g–i). We used a multi-way ANOVA to test the effect of the identity of the immediately preceding and immediately following phrase types on the neural signal ($s = \sum_{t \in P} (\Delta f/f_0)_{\mathrm{denoised}}$). Using Tukey's post hoc comparison and a threshold at $P = 0.05$, we compared the fractions of sequence-correlated ROIs influenced by past phrases, future phrases, or both. This comparison was also carried out separately for ROIs that were active in complex transitions or outside complex transitions (Extended Data Fig. 8h, i).

Testing whether sequence-correlated neurons prefer one or more song contexts. For neurons with significant sequence correlations (one-way ANOVA, described above), we used Tukey's post hoc analysis to determine whether this sequence correlation resulted from a significant single preferred context or significant several preferred contexts (Extended Data Fig. 9). A neuron was declared 'single-context preferring' if the mean signal in only that context was larger than all others (Tukey's $P < 0.001$). A neuron was declared as having preference to more than a single past context if the mean signal following several contexts was larger than another context (Tukey's $P < 0.001$). As the post hoc test uses a subset of the songs, it is weaker than the one-way ANOVA, and some neurons do not show a clear preference to one context or more but still have sequence correlation (grey in Extended Data Fig. 9f).

Maximum fluorescence images for comparing context-dependent signals. For songs that contain a fixed phrase sequence and a variable context element, such as a preceding phrase identity, maximum projection images were created, as above, but using only video frames from the target phrase (for example, the pink phrase in Fig. 2d). Then, the sets of maximum projection images in each context (for example, identity of upstream phrase) were averaged, assigned orthogonal colour maps (for example, red and cyan in Extended Data Fig. 5) and overlaid. Consequentially, regions of the imaging plane that have no sequence preference would be closer to grey scale, whereas ROIs with sequence preference would be coloured. In Extended Data Figs. 5, 9, we used a sigmoidal transform of the colour saturation to amplify the contrast between colour and grey scale without changing the sequence preference information. Additionally, to show that pixels in the ROI are biased towards the same context preference, the above context-averaged maximum projection images were subtracted and pseudo-coloured (insets in Extended Data Fig. 5).

Denoised maximum projection images for comparing context-dependent signals. The maximum projection images described above show the fluorescence signal, including background levels that are typical to single-photon microscopy. To emphasize context-dependent ROIs, we denoised the fluorescence videos using the previously published algorithm CNMFE[49], and created maximum projection images, as above, from the background-subtracted videos (Fig. 4a). The preceding context-preferring ROIs from this estimation algorithm (Fig. 4a) completely overlapped with the manually defined ROIs that were used to extract signal rasters (Fig. 4b). Extended Data Figure 8j replicates Fig. 4a without the de-noising algorithm and shows that the same ROIs report the same context dependence. Supplementary Video 8 shows all the denoised video data that were used to create Fig. 4a.

Label prediction from clustered network states

The signal integral during a target phrase (pink in Extended Data Fig. 10a) was used to create network states—vectors, composed of signals from four jointly recorded ROIs. The averages of the vectors, belonging to the contexts defined by the first upstream (or downstream) phrase label define label-centroids. Then, labels of individual songs were assigned to the nearest neighbouring centroid (Euclidean).

Bootstrapping mutual information in limited song numbers

The neurons in Extended Data Fig. 10a were recorded during 54 songs. This repetition number is too small for estimating the full distribution function of behaviour and network activity states. To overcome this limitation, the mutual information between the network state and the identity of the first upstream (or downstream) phrase was estimated in a bootstrapping permutation process as follows.

We sub-sampled three out of four ROIs in each permutation and converted their signal to binary values by thresholding the signal integral. Next, we reduced the number of phrase labels by merging. Specifically, in Extended Data Fig. 10, the least common label in downstream states was randomly merged with one of the other labels. In the upstream labels, the least common label was merged after a random division of the other four labels, to form two groups of two.

The mutual information measures were then calculated for each of the 48 possible state spaces and divided by the entropy of the behaviour state, leading to the scatter shown in Extended Data Fig. 10b. The margin of error was estimated from the standard deviation. The 0.95 quantile level of the null hypothesis was created by randomly shuffling each variable to create 1,000 surrogate datasets and repeating the measures. The shuffled set was used to create a sample distribution and to calculate the significance of the differences in Extended Data Fig. 10b using a *z*-test with the sample mean and standard deviation.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data can be found at figshare (https://figshare.com/) with https://doi.org/10.6084/m9.figshare.12006657. Source data are provided with this paper.

Code availability

All custom-made code in this manuscript is publicly available in Github repositories (https://github.com/gardner-lab/FinchScope; https://github.com/gardner-lab/video-capture; https://github.com/gardner-lab/FinchScope/tree/master/Analysis%20Pipeline/extract-media; https://github.com/yardencsGitHub/BirdSongBout/tree/master/helpers/GUI; https://github.com/yardencsGitHub/tweetynet; and https://github.com/jmarkow/pst).

45. Stokes, T. M., Leonard, C. M. & Nottebohm, F. The telencephalon, diencephalon, and mesencephalon of the canary, *Serinus canaria*, in stereotaxic coordinates. *J. Comp. Neurol.* **156**, 337–374 (1974).
46. Liberti, W. A., III et al. Unstable neurons underlie a stable learned behavior. *Nat. Neurosci.* **19**, 1665–1671 (2016).
47. Wild, J. M., Williams, M. N., Howie, G. J. & Mooney, R. Calcium-binding proteins define interneurons in HVC of the zebra finch (*Taeniopygia guttata*). *J. Comp. Neurol.* **483**, 76–90 (2005).
48. Wohlgemuth, M. J., Sober, S. J. & Brainard, M. S. Linked control of syllable sequence and phonology in birdsong. *J. Neurosci.* **30**, 12936–12949 (2010).
49. Zhou, P. et al. Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *eLife* **7**, e28728 (2018).
50. Howell, D. C. *Statistical Methods for Psychology* (Cengage Learning, 2009).

Acknowledgements This study was supported by NIH grants R01NS089679, R01NS104925, R24NS098536 (T.J.G.) and R24HL123828, U01TR001810 (D.N.K.) We thank J. Markowitz, I. Davison, and J. Gavornik for discussions and comments on this manuscript, and Nvidia Corporation for a technology grant (Y.C.).

Author contributions Y.C. and T.J.G. conceived and designed the study. W.A.L.III designed miniaturized microscopes and tether commutators and consulted on surgical procedures. L.N.P. created the video acquisition software. D.C.L. and D.N.K. produced lentivirus. Y.C. and J.S. designed surgical procedures. Y.C., J.S., and D.S. performed animal surgeries. Y.C. and D.P.L. built the experimental setup. Y.C. and J.S. gathered the data. Y.C. and D.S. performed histology and immunohistochemistry. Y.C. designed and wrote the machine-learning audio segmentation and annotation algorithm. Y.C. analysed the data. Y.C., W.A.L.III, L.N.P., and T.J.G. wrote the manuscript.

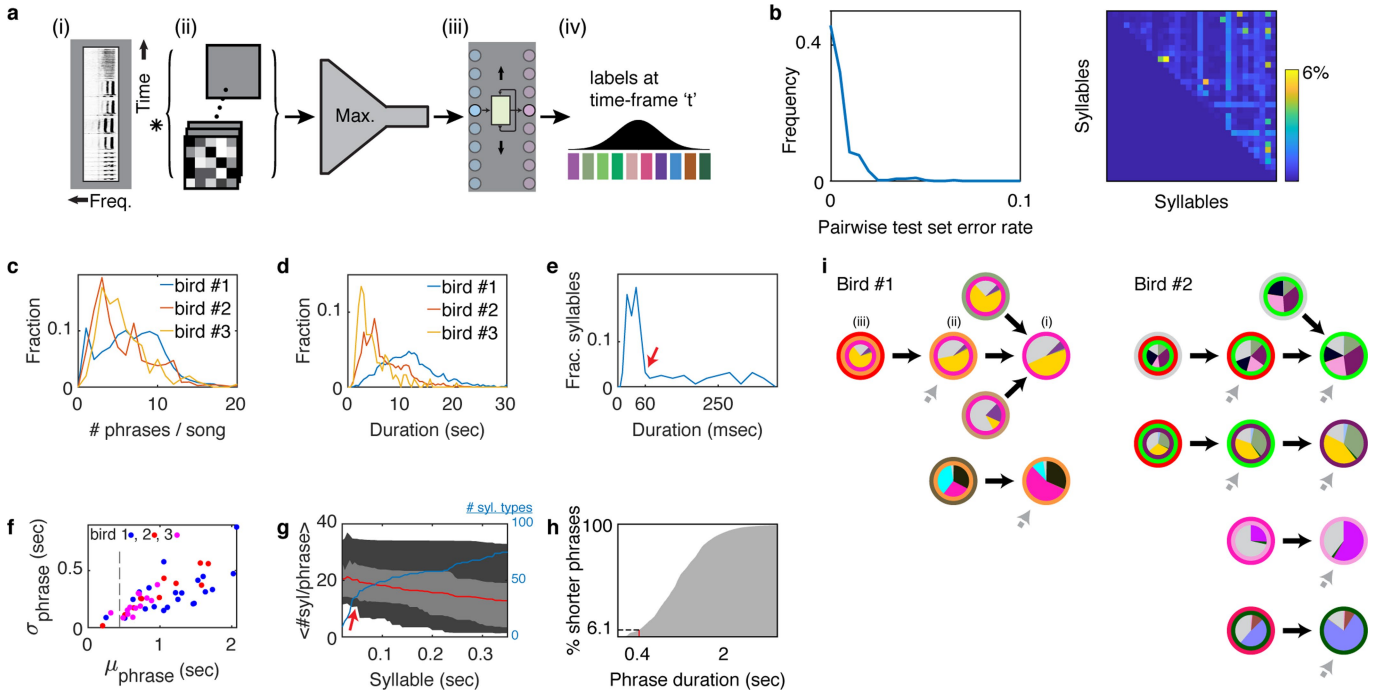
Competing interests The authors declare no competing interests.

Additional information

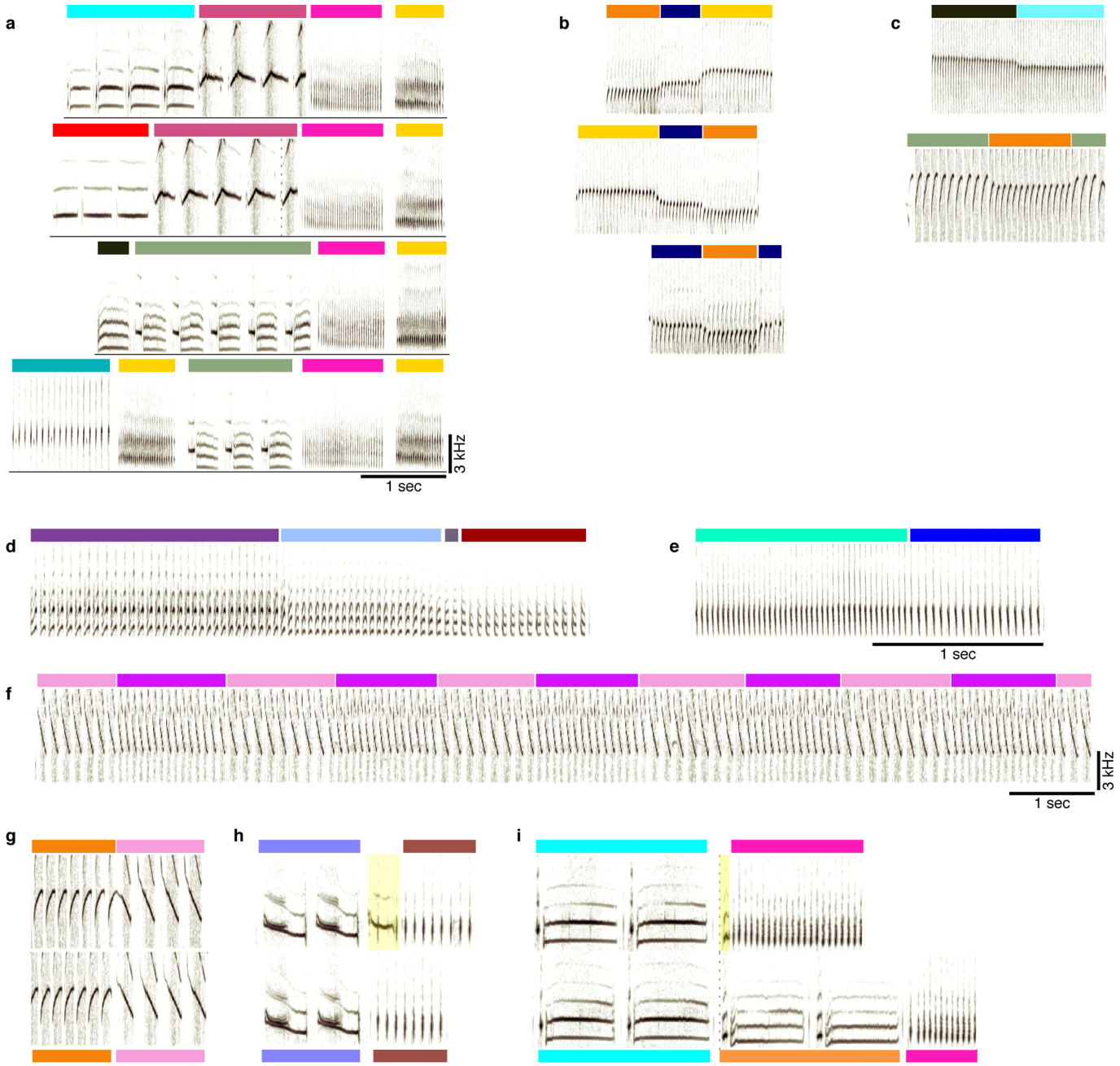
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2397-3.

Correspondence and requests for materials should be addressed to Y.C. or T.J.G.

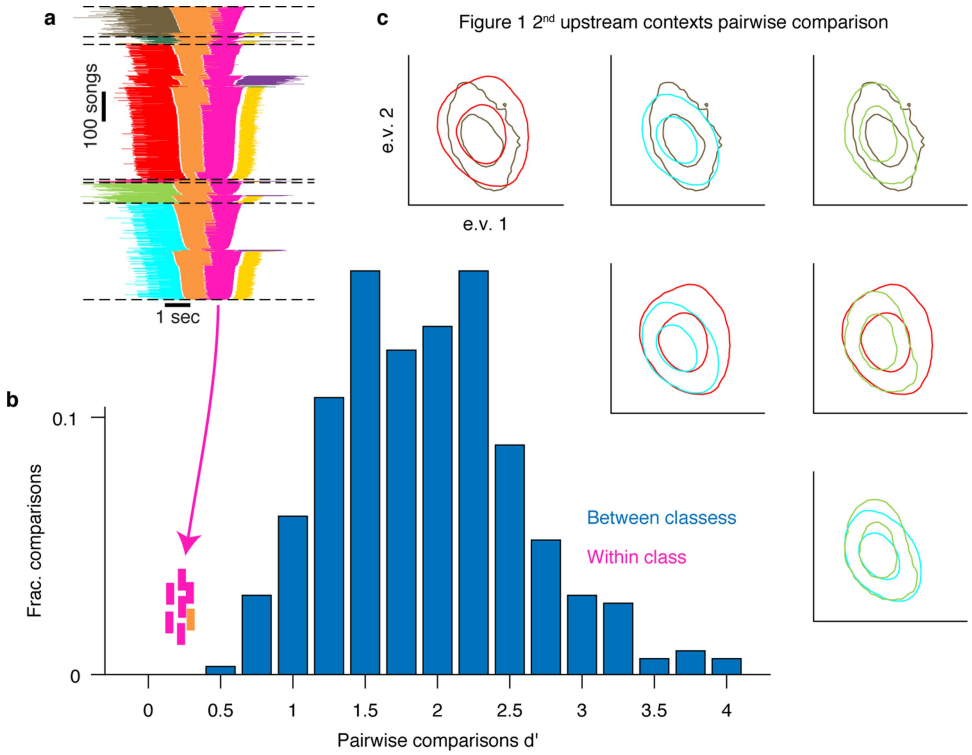
Peer review information *Nature* thanks Jesse Goldberg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



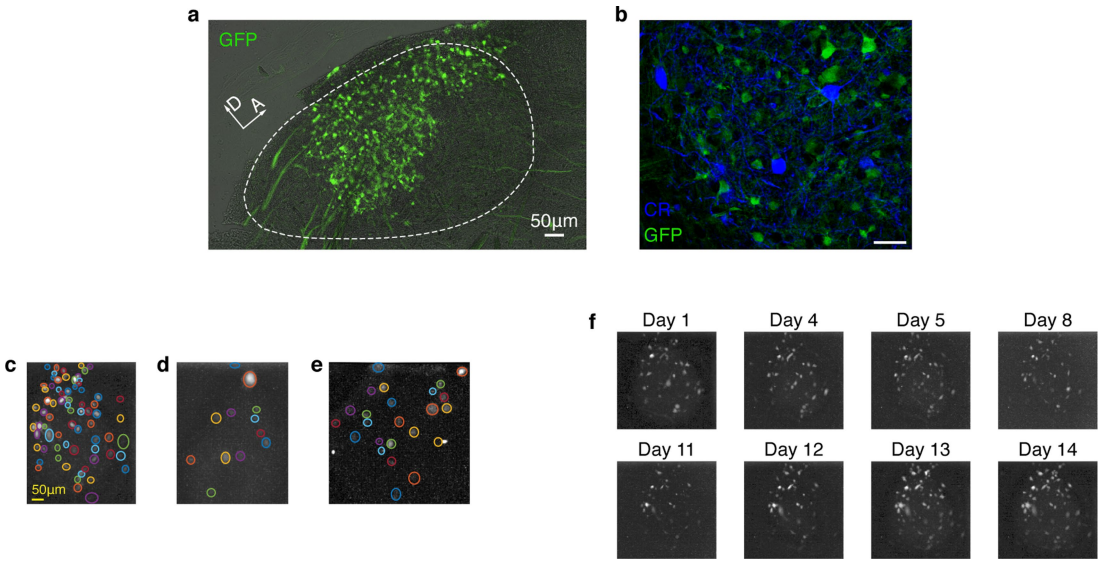
Extended Data Fig. 1 | Canary song annotation and sequence statistics. **a**, Architecture of syllable segmentation and annotation machine learning algorithm. (i) A spectrogram is fed into the algorithm as a 2D matrix in segments of 1 s. (ii) Convolutional and max-pooling layers learn local spectral and temporal filters. (iii) Bidirectional recurrent LSTM layer learns temporal sequencing features. (iv) Projection onto syllable classes assigns a probability for each 2.7-ms time bin and syllable. **b**, After manual proofreading (see Methods), a support vector machine classifier was used to assess the pairwise confusion between all syllable classes of bird 1 (see Methods). The test set confusion matrix (right) and its histogram (left) show that in rare cases the error exceeded 1% and at most reached 6%. As the higher values occurred only in phrases with 10 s of syllables, this metric guarantees that most of the syllables in every phrase cannot be confused as belonging to another syllable class. Accordingly, the possibility of making a mistake in identifying a phrase type is negligible. **c**, Number of phrases per song for the three birds used in this study. **d**, Song durations for the three birds. **e**, Mean syllable durations for 85 syllable classes from three birds. Red arrow marks the duration below which all trill types have more than ten repetitions on average. **f**, Relation between phrase class mean duration (*x* axis) and standard deviation (*y* axis). Syllable classes (dots) of three birds are coloured according to bird number. Dashed line marks 450 ms (upper limit for the decay time constant of GCaMP6f). **g**, Range of mean number of syllables per phrase (*y* axis) for all syllable types with mean duration shorter than the *x*-axis value. Red line is the median, light grey marks the 25% and 75% quantiles and dark grey marks the 5% and 95% quantiles (blue line marks the number of syllable types contributing to these statistics). The red arrow matches the arrow in **e**. **h**, Cumulative histogram of trill phrase durations. **i**, All complex phrase transitions with second-order or higher dependence on song history context (for birds 1 and 2). For each phrase type that precedes a complex transition, the context dependence is visualized by a PST (see Methods). Transition outcome probabilities are marked by pie charts at the centre of each node. The song context (phrase sequence) that leads to the transition is marked by concentric circles, the innermost being the phrase type that preceded the transition. Nodes are connected to indicate the sequences in which they are added in the search for longer Markov chains that describe context dependence (for example, i–iii for first- to third-order Markov chains). Grey arrows indicate additional incoming links that are omitted for simplicity.



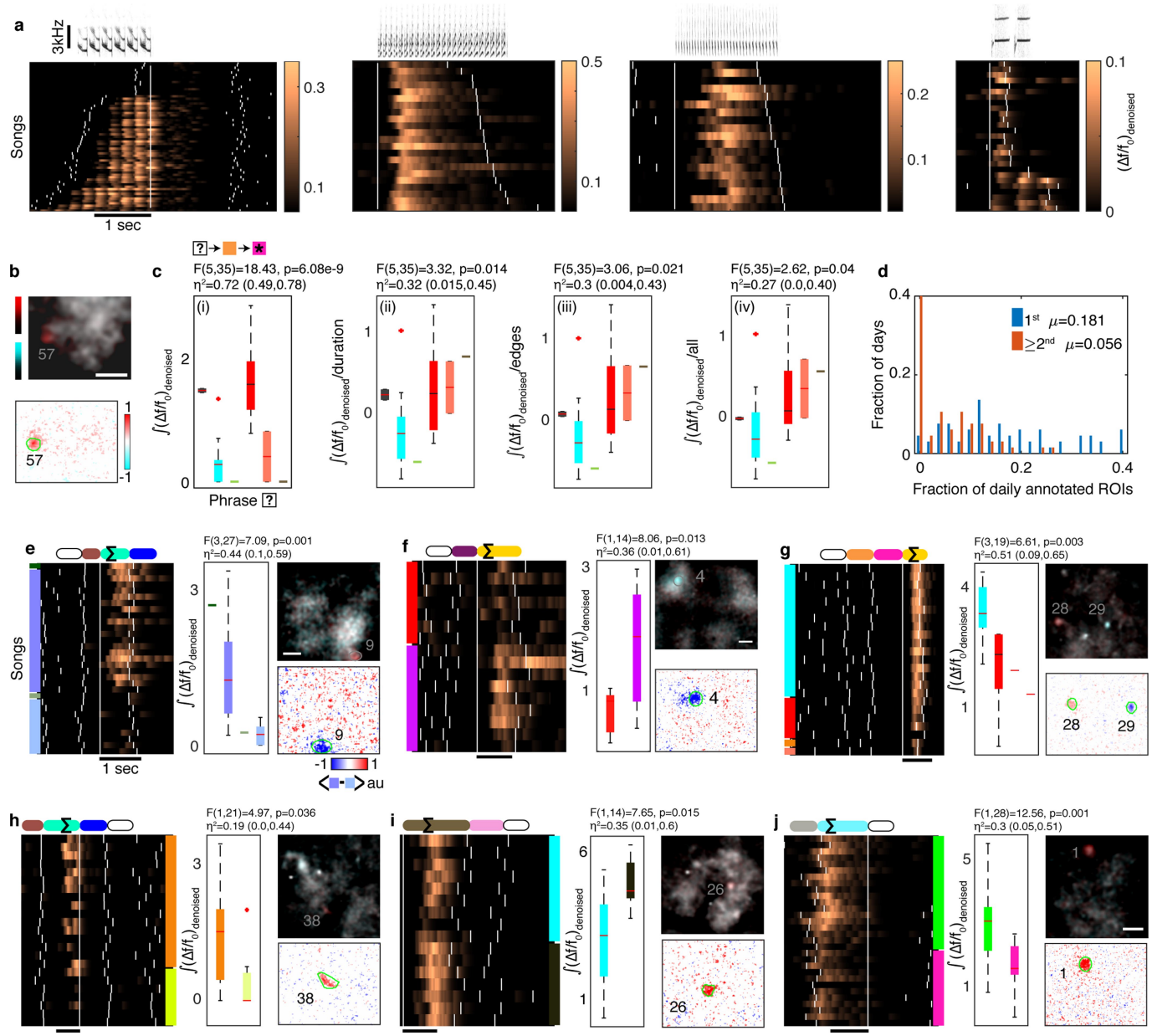
Extended Data Fig. 2 | Examples of canary song phrase sequences, rare inter-phrase gaps, and aberrant syllables. a, Additional spectrograms of phrase sequences (colours above the spectrograms indicate phrase identity) that lead to a repeating pair of phrases (pink and yellow). **b**, Examples of flexible phrase sequencing comprising pitch changes (from bird 3). **c**, Examples of phrase transitions with a pitch change from bird 2. **d**–**f**, Phrase sequences showing changes in spectral and temporal parameters. **d**, Bird 1 changes from up sweep (purple) to down sweep (dark red) through intermediate phrases of intermediate acoustic structure. **e**, Bird 1 shows a change in inter-syllable gaps. **f**, Bird 2 shows changes in pitch sweep rate. **g**, Top and bottom sonograms compare the same phrase transitions where the inter-phrase gap varies. **h**, **i**, The top sonogram includes a rare vocalization at the beginning of the second phrase (highlighted) that, in **i**, resembles the onset of an orange phrase type.



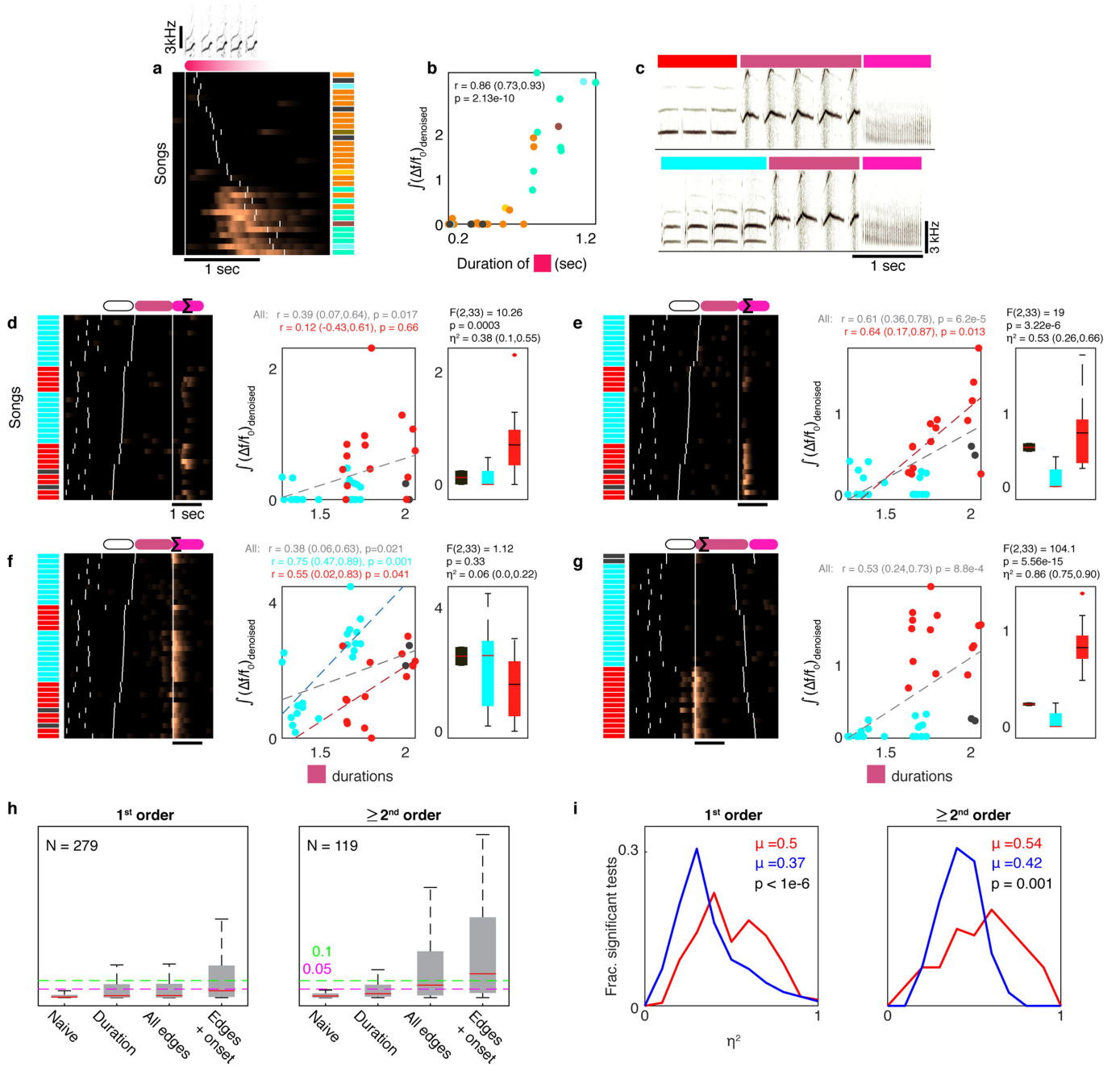
Extended Data Fig. 3 | An example in which the context-dependence of syllable acoustics before complex transitions is too small for clear distinction. a, Same as Fig. 1b. A summary of all phrase sequences that contain a common transition reveals that the choice of what to sing after the pink phrase depends on the phrases that were produced earlier. Lines represent phrase identity and duration. Song sequences are stacked (vertical axis) sorted by the identity of the first phrase, the identity of the last phrase, and then the duration of the centre phrases. **b**, The discriminability (d', x axis) measures the acoustic distance between pairs of syllable classes in units of the within-class standard deviation (see Methods). Bars show the histogram across all pairs of syllables identified by human observers (see Methods), corresponding to about 99% or more identification success (Extended Data Fig. 1b). The pink ticks mark the d' values for six within-class comparison of the main four contexts in **a**. The orange tick marks the d' for another context comparison in a different syllable that precedes a complex transition for this bird. **c**, The pairwise comparison of distributions matching the pink ticks in **b**. Each inset shows overlays of two distributions marked by contours at the 0.1 and 0.5 values of the peak and coloured according to context in **a**. The distributions are projected onto the two leading principle components of the acoustic features (see Methods, in the space defined by eight acoustic features[48]). While some of these distributions are statistically distinct, they allow for only about 70% context identification success in the most distinct case.



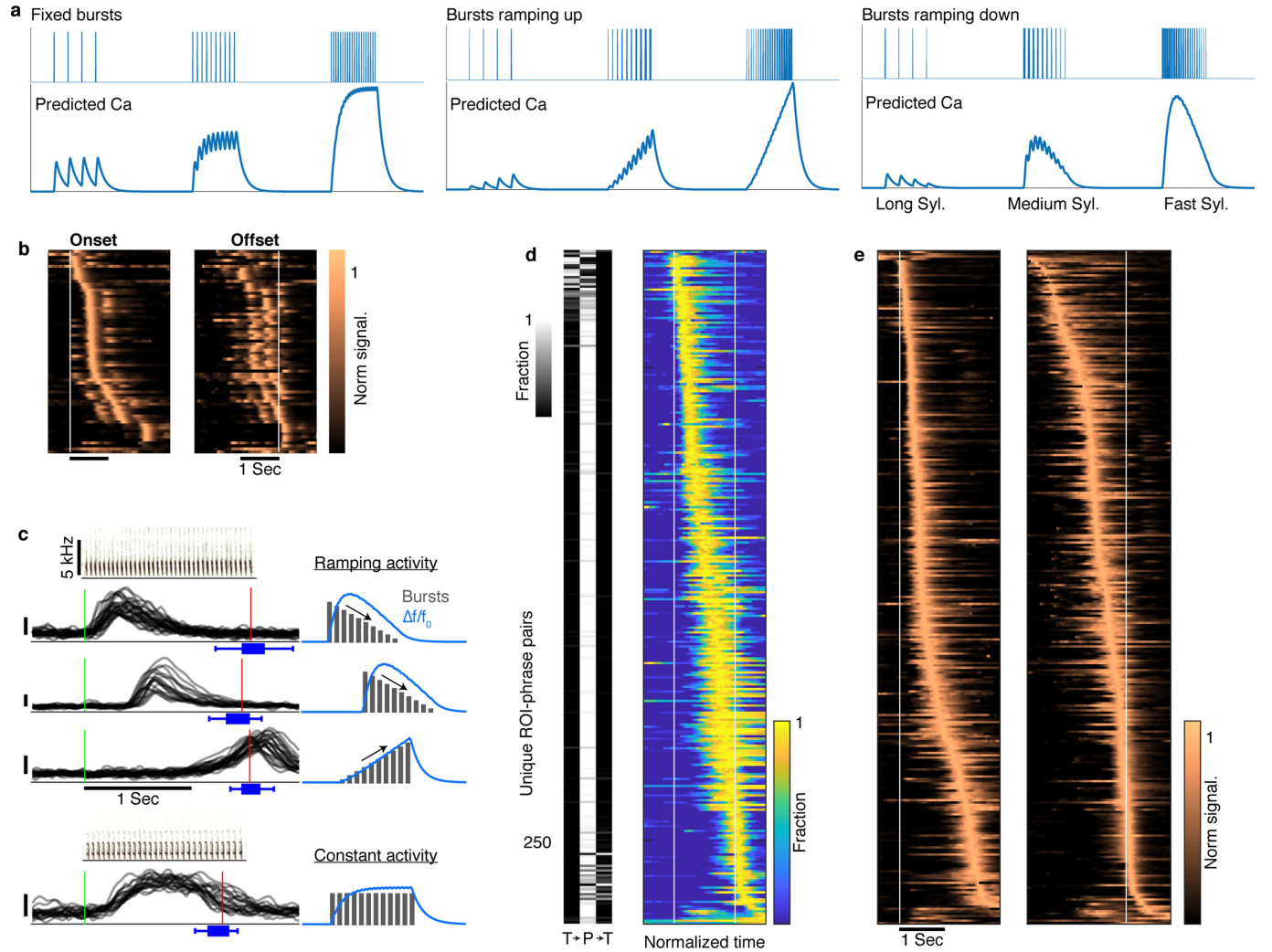
Extended Data Fig. 4 | Calcium indicator is expressed exclusively in HVC excitatory neurons and imaged in annotated ROIs. a, Sagittal slice of HVC showing GCaMP-expressing PNs (experiment repeated in five birds with similar results). **b**, We observed no overlap between transduced GCaMP6f-expressing neurons and neurons stained for the inhibitory neurons markers calretinin (CR), calbindin, and parvalbumin (calretinin stain shown, staining experiment repeated six times for each marker with similar results). **c**–**e**, Examples of daily ROI annotation in three birds (1–3). Coloured circles mark different ROIs, manually annotated on maximum fluorescence projection images on an exemplary day (see Methods). **f**, Maximum fluorescence images (from bird 1; see Methods) revealing fluorescence sources, including sparsely active cells, in the imaging window across multiple days.



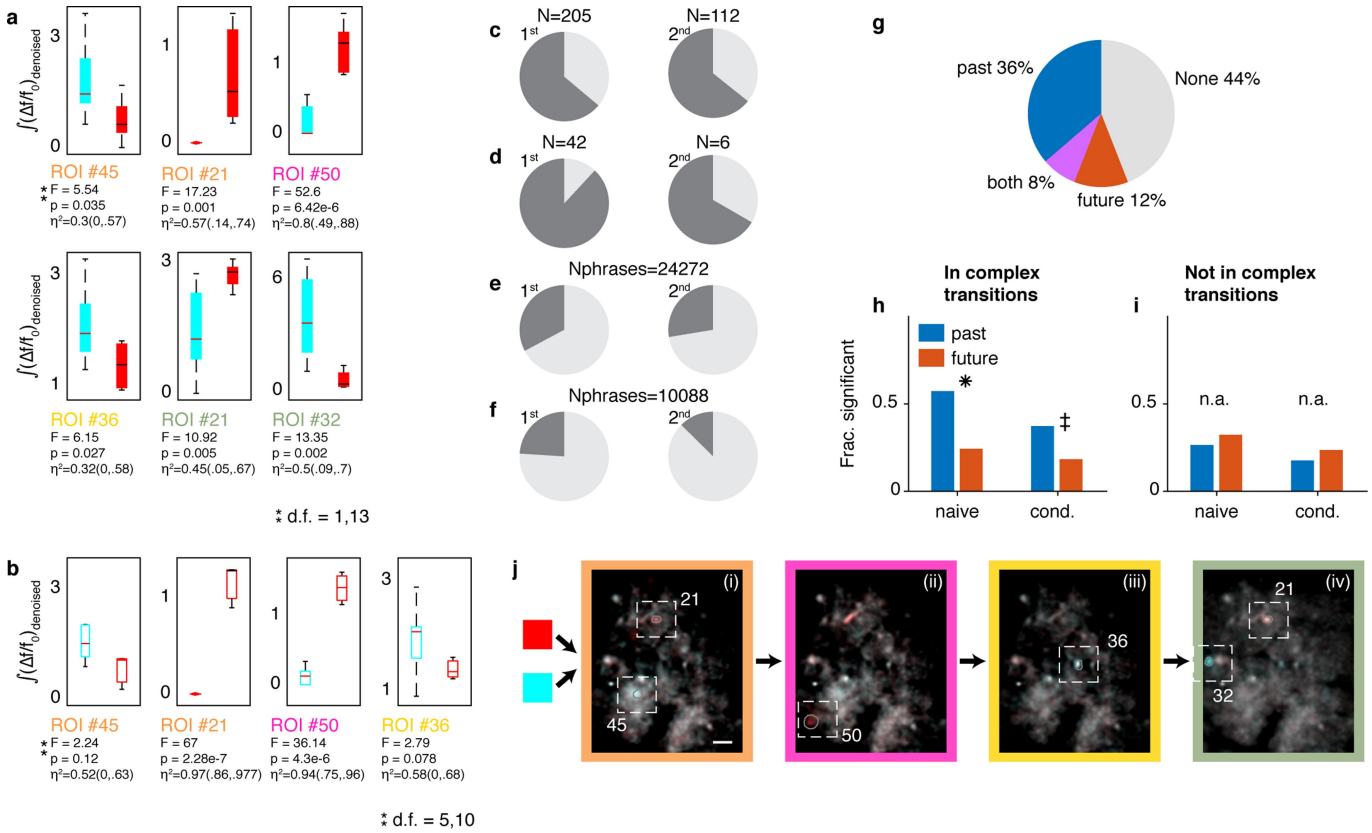
Extended Data Fig. 5 | Syllable and phrase-sequence-correlated ROIs from three birds. a, Sonograms above rasters from four ROIs from three birds. White ticks indicate phrase onsets. The fluorescent calcium indicator is able to resolve individual long syllables. **b**, Top, average maximum fluorescence images during the pink phrase in Fig. 2d (compare the two most common contexts in orthogonal colours (red and cyan)). Scale bar, is 50 μm. Bottom, difference of the overlaid images. ROI outlined in green. **c**, (i) One-way ANOVA (F, P, η^2 and its 95% CI) tests the effect of contexts (x axis, second preceding phrase type in $n = 41$ sequences) on the signal (y axis) during the target phrase (marked by star) in Fig. 2d. Lines, boxes, whiskers, and plus symbols show the median, first and third quartiles, full range, and outliers. (ii–iv) ANOVA tests carried out using the residuals from the signal after removing the cumulative linear dependence on the duration of the target phrase, the relative timing of onset and offset edges of two fixed phrases, and the absolute onset time of the target phrase in each rendition. Colours correspond to phrases in Fig. 2d. **d**, Fractions of daily annotated ROIs showing sequence correlation in all three birds. Each ROI can be counted only once per order. This estimate includes sparsely active ROIs. **e**–**j**, Activity during a target phrase (marked by Σ) is strongly related to non-adjacent phrase identities (empty lozenges in colour-coded phrase sequence). Songs are arranged by the phrase sequence context (left or right colour patches for past and future phrase types, respectively). White ticks indicate phrase onsets. Box plots and contrast images as defined in **b**, **c**. $n = 31, 16, 23, 23, 16$ and 30 songs contribute to **e**–**j**, respectively. **e**, **f**, Similar to main Fig. 2d, $(\Delta f/f_0)_{denoised}$ from ROIs with second-order upstream sequence (colour coded) from two more birds. **g**, Third-order upstream relation. **h**, **i**, Second-order downstream relations. **j**, First-order downstream relation from another bird.



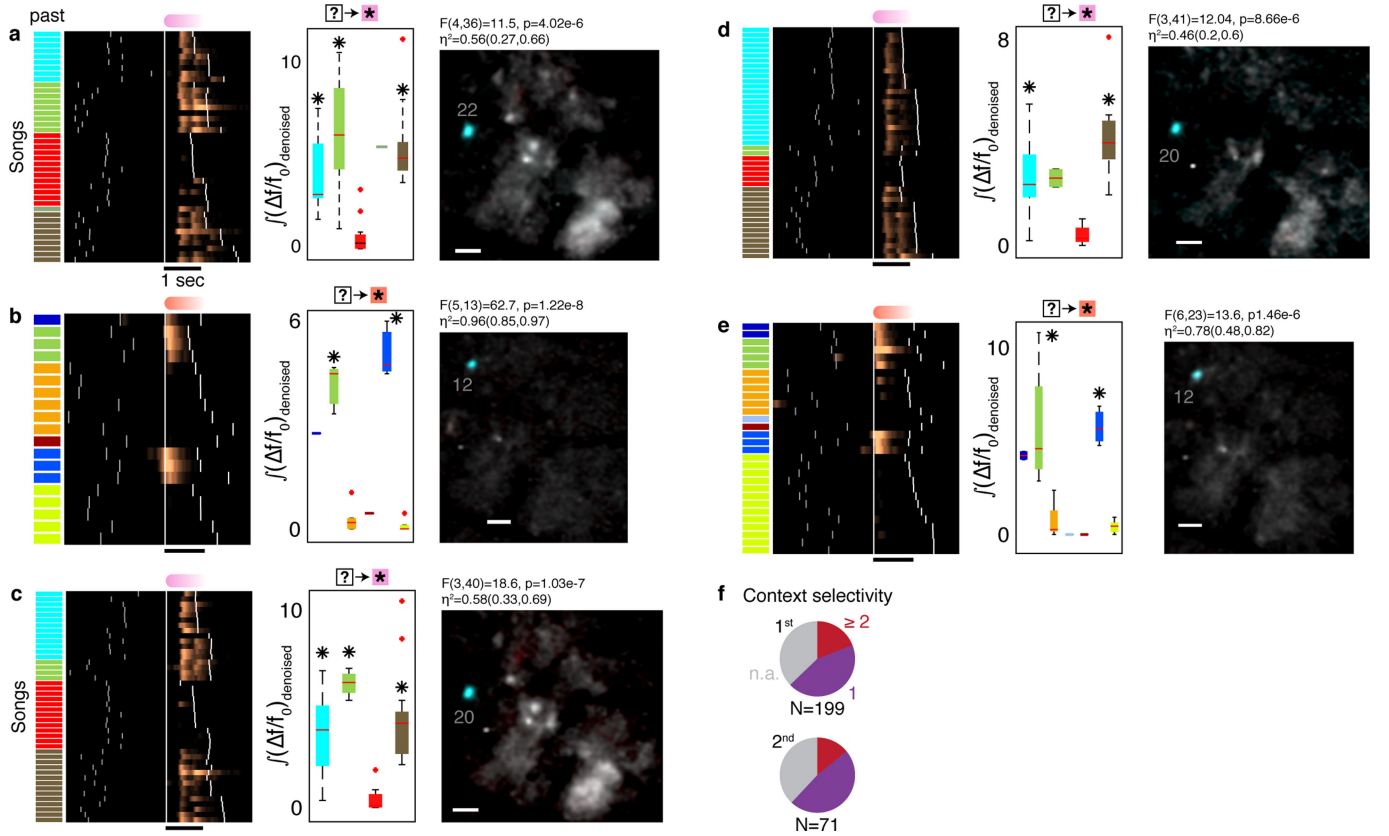
Extended Data Fig. 6 | Durations and onset times of phrases also correlate with their sequence, but cannot fully account for HVC activity. **a**, $(\Delta f/f_0)_{denoised}$ signal traces (ROI 18, bird 3) during one phrase type (red) arranged by phrase duration. Coloured barcode annotates the final phrase in the sequence. **b**, The signal correlates to the red phrase's duration (r (95% CI), P: Two-sided Pearson's test for $n = 32$ songs). Colours match barcode in **a**. **c**, Sonograms of two phrase sequences. **d**–**g**, ROI signals during $n = 36$ sequences containing the last two phrases in **c** have various relations to the duration of the middle (purple) phrase (middle; scatter plots as in **b**, dashed lines indicate significant correlations) and the identity of the first phrase (right; colours, one-way ANOVA (F, P, η^2 (95% CI)) tests the effect on the signal Σ. Whiskers, boxes, and lines show full range, first and third quartiles, and medians, respectively). **d**, Signal correlation with phrase duration is completely entangled with the signal's sequence preference and does not apply in separate preceding contexts (red, $P > 0.5$). **e**, Signal correlation with phrase duration is influenced by the signal's sequence preference but also exists in the preferred sequence context separately (red). **f**, Signal duration correlation is observed within each single preceding context separately, but the correlation reduces across all songs. **g**, Similar to **a**, but the signal is in the second phrase, not the third. **h**, Distributions of one-way ANOVA P values (y axis; whiskers, boxes, and red lines show full range, first and third quartiles, and medians, respectively) relating phrase identity and signal for adjacent phrases ($n = 279$ independent first-order tests, left) and non-adjacent phrases ($n = 119$ independent second- or higher-order tests, right). Tests were also done on residuals of signals, after discounting the following variables: variance explained by the target phrase duration, the timing of all phrase edges in the test sequence, and the time-in-song (x axis, effects accumulated left to right by multivariate linear regression; see Methods). Coloured dashed lines mark $P = 0.05$ and 0.1. **i**, Effect size (η^2 denotes fraction of variance accounted for by the signals' context dependence) of past (red) and future (blue) one-way ANOVA tests for first-order (left, $N = 279$ tests) and second- or higher-order (right, $n = 119$) correlations. The difference in the mean value (μ) is tested using one-sided bootstrap shuffles (P values, see Methods).



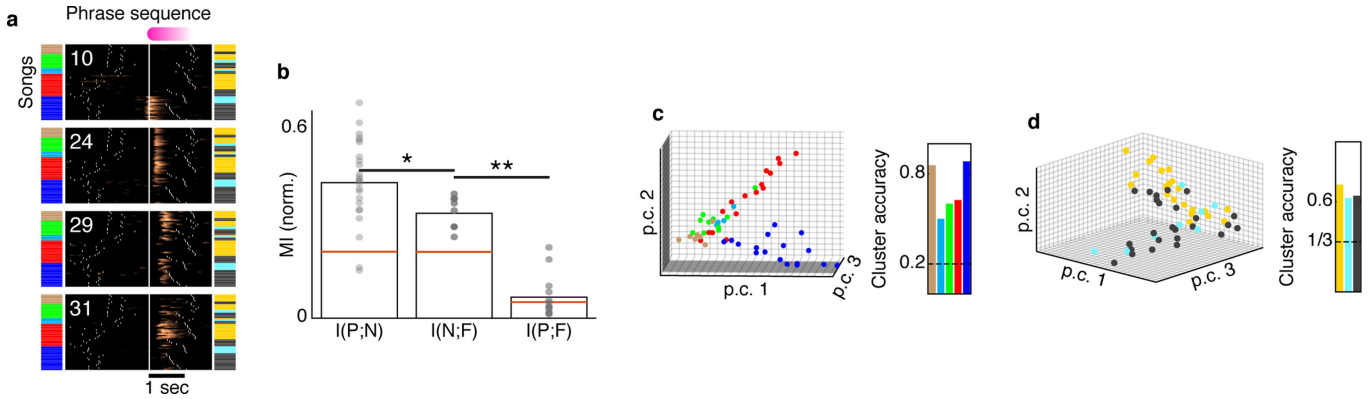
Extended Data Fig. 7 | Signal shape and onset time of sequence-correlated HVC neuron activity reflect within-phrase timing. a, Simulation of calcium indicator (GCaMP6f) fluorescence corresponding to syllable-locked spike bursts in HVC PNs. Syllable-locked spike bursts are convolved with the indicator's kernel (see Methods) to estimate the expected signal when the number of spikes per burst is constant (left), ramps up (middle), or ramps down (right) linearly with the syllable number. The simulation assumes one burst per syllable in time spacing (x axis) that matches long canary syllables (400–500ms), medium-length syllables (100 ms) and short syllables (50 ms). **b**, Complementing Fig. 3a, average context-sensitive activity in phrases with long syllables reveals syllable-locked peaks aligned to phrase onsets (left) or offsets (right, same row order as left) that change in magnitude across the phrase. **c**, Signal shape and onset timing have properties of within-phrase timing codes. Example raw $\Delta f/f_0$ signals (y axis, 0.1 marked by vertical bar) of four ROIs aligned to the onset of specific phrase types (green line). Sonograms show the repeating syllables. Red lines and blue box plots show the median, range, and quartiles of the phrase offset timing. The signal shapes resemble the expected fluorescence of the calcium indicator elicited by syllable-locked ramping (sketches, top three) or constant activity (bottom). **d**, Left, barcodes show the fraction of signal onsets found in the preceding transition, within the phrase, and in the following transition (T→P→T, see Methods). Rows correspond to the phrases in Fig. 3a. Right, rows show the average signal state occupancy estimated from HMMs fitted to the single-trial data used for Fig. 3a. The resulting traces are time-warped to fixed phrase edges (white lines). **e**, Single-trial data from Fig. 3a aligned to phrase onsets (left) and offsets (right) and averaged in real time. The resulting traces are ordered by peak location (separately in left and right rasters).



Extended Data Fig. 8 | Context-sensitive signals aggregate in complex transitions and preferentially encode past transitions. a, Distribution of signal integrals (y axis; whiskers show full range, boxes show first and third quartiles, and lines show medians) for ROIs in Fig. 4a. Text label is colour coded by phrase type in i–iv. F numbers, P values, and η^2 (95% CI) for one-way ANOVA relating history (x axis) and signal (y axis) in $n = 15$ song sequences. **b**, ROIs in **a** retain their song-context bias for songs that terminate at end of the third phrase rather than continuing. Box plots repeat the ANOVA tests in **a** for $n = 16$ songs in which the last phrase is replaced by the end of the song. **c**–**f**, Dark grey slices indicate the fraction of correlations that occur in complex behavioural transitions. **c**, **d**, Data from Fig. 4c separated into the two birds. **e**, **f**, The fraction in **c**, **d** expected by the null hypothesis of correlations distributing by the frequency of each phrase type among $N_{phrases}$ phrases in the dataset. **g**, In sequence-correlated ROIs, multi-way ANOVA is used to separate the effects of the preceding and following phrase types on the signal (see Methods). Pie chart shows the percentage of sequence-correlated ROIs that were significantly influenced by the past, future, or both phrase identities among $n = 336$ significant ANOVA tests. **h**, Restricting analysis to complex transitions, more ROIs correlated with the preceding phrase type (blue) than with the following one (red). This is true in both naive signal values (left, $n = 185$ tests) and after we removed dependencies on phrase durations and time-in-song (right, $n = 185$). One-sided binomial z-test: *proportion difference 0.33 ± 0.09, $Z = 6.45$, $P = 5.5 \times 10^{-11}$; ‡proportion difference 0.19 ± 0.09, $Z = 4.05$, $P = 2 \times 10^{-5}$. **i**, Restricting the analysis to phrase types that are not in complex transitions ($n = 136$ ANOVA tests) reveals more ROIs correlated with the future phrase type, but the difference is not significant (left, right, n.a.: one-sided binomial z-test, $P = 0.14$, 0.11). **j**, Fig. 4a showed maximum projection images, calculated with denoised videos (see Methods). The algorithm CNMF-E[49] involves estimating the source ROI shapes, de-convolving spike times and estimating the background noise. Here, recreating the maximum projection images with the original fluorescence videos shows the background as well, but the preceding-context-sensitive neurons remain the same. Namely, the same ROI footprints annotated in i–iv show the colour bias (cyan or red) that indicates coding of the past phrase with the same colour.



Extended Data Fig. 9 | ROIs that reflect several preceding song contexts. **a**, **b**, ROIs that are active in multiple preceding contexts. $(\Delta f/f_0)_{denoised}$ traces are aligned to a specific phrase onset, arranged by identity of the preceding phrase (colour barcode). White ticks indicate phrase onsets. Box plot shows distributions of $(\Delta f/f_0)_{denoised}$ integrals (*y* axis, summation in the phrase marked by star) for various song contexts (*x* axis). *F* number, *P* value, and effect size (η^2 (95% CI)) show the significance of separation by song context (one-way ANOVA). Asterisks mark contexts that lead to larger mean activity compared to another context (Tukey's multiple comparisons; $n = 41$ songs and $P = 0.01, 7.5 \times 10^{-6}, 5.6 \times 10^{-5}$ in **a**; $n = 19$, $P = 8.8 \times 10^{-7}, 8.15 \times 10^{-8}$ in **b**). Average maximum projection images (see Methods) during the aligned phrase compare the song contexts that lead to significantly higher activity with the other contexts in orthogonal colours (cyan and red for high and low activity, respectively). Scale bar, 50 µm. **c**–**e**, Neurons with similar context preference to those in **a** and **b** on adjacent days. Tukey's multiple comparisons: $n = 44$, $P = 0.001, 4.08 \times 10^{-6}, 1.3 \times 10^{-6}$ in **c**; $n = 45$, $P = 0.0016, 2.85 \times 10^{-6}$ in **d**; $n = 30$, $P = 0.0002, 0.0001$ in **e**. **f**, Fraction of ROIs with selectivity for one context (purple) or multiple contexts (red) identified using Tukey's post hoc multiple comparisons (see Methods). Grey slices (n.a.) mark context-sensitive ROIs for which the post hoc analysis did not isolate a specific context with a larger mean signal. Top (bottom) pie shows selectivity for first (second) preceding phrases.



Extended Data Fig. 10 | HVC neurons can be tuned to complementary preceding contexts. a, Four jointly recorded ROIs exhibit complementary context selectivity. Colour bars indicate phrase identities preceding and following a fixed phrase (pink). For each ROI (rasters), $(\Delta f/f_0)_{\text{denoised}}$ traces are aligned to the onset of the pink phrase (x axis) arranged by the identity of the preceding phrase, by the identity of the following phrase, and finally by the duration of the pink phrase. **b**, For the example in **a**, normalized mutual information between the identity of past (P) and future (F) phrase types is significantly smaller than the information held by the network states about the past and future contexts (left bars; N is the activity of the four ROIs). Dots, bars, and red lines mark bootstrap assessment shuffles, their means, and the 95% level of the mean in shuffled data (see Methods). *Difference is 0.09 ± 0.03, $Z = 4.3$, $P = 7.3 \times 10^{-6}$; **difference is 0.26 ± 0.02, $Z = 8.9$, $P < 1 \times 10^{-15}$, bootstrapped one-sided z-test. **c**, Signal integrals from the four ROIs in **a** are plotted for each song (dots, $n = 54$ songs) on the three most informative principle components. Dots are coloured by the identity of the preceding phrase. Clustering accuracy measures the 'leave-one-out' label prediction for each preceding phrase (true positive), calculated by assigning each dot to the nearest centroid (L_2). Dashed line marks chance level. **d**, As in **c** but for the first following phrase.

nature research

Corresponding author(s): Yarden Cohen

Last updated by author(s): Mar 19, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	VOS games' Boom Recorder was used for audio recording from cage microphones. In-house developed software was used for audio and video acquisition in imaging experiments (https://github.com/gardner-lab/video-capture). Histology images were taken with Nikon's 'Elements AR' v4.51.01
Data analysis	Data was analyzed by in-house developed Matlab R2009, R2016b,R2017b (Mathworks) and python (3.6.3) programs. Github links provided in the methods description.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets are available from the corresponding author on request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample sizes are similar to sample sizes used in the field. No statistical methods were used to determine sample size.
Data exclusions	We did not exclude any animal for data analysis. We excluded data from the very rare occasions in which video files were corrupted because of tethering malfunctions.
Replication	We performed recordings from multiple animals to confirm reproducibility. Replications were successful.
Randomization	Our study did not include experimental groups and did not require randomization.
Blinding	Our study did not include experimental groups and did not require blinding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Immunohistochemistry:
Anti-Calbindin(SWANT CB38, rabbit, lot # 9.03, 1:4000), Anti-Calretinin(SWANT 7697, rabbit, lot # 1893-0114,1:15000), Anti-Parvalbumin (SWANT PV27, rabbit, lot # 2014, 1:1000)

Validation

All antibodies are used in a large number of publications and do not require additional validation.

References in manufacturer's website:
Anti-Calbindin:
Airaksinen M.S., et al, (1997), PNAS 94(4) : 1488-1493
Anti-Calretinin:
1. Schwaller B., Buchwald P., Blümcke I., Celio M.R. and Hunziker W. (1994) Characterization of a polyclonal antiserum against the purified human recombinant calcium-binding protein calretinin. Cell Calcium 14: 639-648.
2. Schiffmann S.N. et al (1999) Impaired motor coordination and Purkinje cell excitability in mice lacking calretinin. PNAS, 96: 5257-5262.
3. Gotzos V., Vogt P. and M.R. Celio (1995) Calretinin is a selective marker for malignant pleural mesotheliomas of the epithelial type. Pathol. Res. Pract. 192:137-147.
4. Doglioni, C. et al. (1996) Calretinin: a novel immunocytochemical marker for mesothelioma. Am. J. Surg. Pathol. 20:1037-1046.
Anti-Parvalbumin:
1. Kretsinger R.H. (1981) Neurosci. Res. Progr. Bull. 19/8, MIT-Press
2. Celio M.R., Heizmann C.W. (1981) Nature 293: 300-302
3. Celio M.R., Heizmann C.W. (1982) Nature 297:504-506
4. Schwaller B., et al. (1999) Am. J. Physiol. 276. C395-403

5. Filice F, Celio M.R., Szabolcsi V. (2017) JCN

References in songbird literature:
1. Wild, J. M., Williams, M. N., Howie, G. J. & Mooney, R. Calcium-binding proteins define interneurons in HVC of the zebra finch (Taeniopygia guttata). Journal of Comparative Neurology 483, 76–90 (2005).
2. Scotto-Lomassese, S., Rochefort, C., Nshdejan, A. & Scharff, C. HVC interneurons are not renewed in adult male zebra finches. European Journal of Neuroscience 25, 1663–1668.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Domestic canaries (Serinus Canaria), American Singer strain, males, older than 1 year.
Wild animals	The study did not involve wild animals
Field-collected samples	The study did not involved samples collected from the field
Ethics oversight	All procedures were approved by the Institutional Animal Care and Use Committee of Boston University (protocol numbers 14-028 and 14-029) with accreditation from the Association for Assessment and Accreditation of Laboratory Animal Care International.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Feedback generates a second receptive field in neurons of the visual cortex

https://doi.org/10.1038/s41586-020-2319-4

Received: 21 June 2019

Accepted: 10 March 2020

Published online: 20 May 2020

Check for updates

Andreas J. Keller[1,2]✉, Morgane M. Roth[1,2] & Massimo Scanziani[1,2]✉

Animals sense the environment through pathways that link sensory organs to the brain. In the visual system, these feedforward pathways define the classical feedforward receptive field (ffRF), the area in space in which visual stimuli excite a neuron[1]. The visual system also uses visual context—the visual scene surrounding a stimulus—to predict the content of the stimulus[2], and accordingly, neurons have been identified that are excited by stimuli outside their ffRF[3–8]. However, the mechanisms that generate excitation to stimuli outside the ffRF are unclear. Here we show that feedback projections onto excitatory neurons in the mouse primary visual cortex generate a second receptive field that is driven by stimuli outside the ffRF. The stimulation of this feedback receptive field (fbRF) elicits responses that are slower and are delayed in comparison with those resulting from the stimulation of the ffRF. These responses are preferentially reduced by anaesthesia and by silencing higher visual areas. Feedback inputs from higher visual areas have scattered receptive fields relative to their putative targets in the primary visual cortex, which enables the generation of the fbRF. Neurons with fbRFs are located in cortical layers that receive strong feedback projections and are absent in the main input layer, which is consistent with a laminar processing hierarchy. The observation that large, uniform stimuli—which cover both the fbRF and the ffRF—suppress these responses indicates that the fbRF and the ffRF are mutually antagonistic. Whereas somatostatin-expressing inhibitory neurons are driven by these large stimuli, inhibitory neurons that express parvalbumin and vasoactive intestinal peptide have mutually antagonistic fbRF and ffRF, similar to excitatory neurons. Feedback projections may therefore enable neurons to use context to estimate information that is missing from the ffRF and to report differences in stimulus features across visual space, regardless of whether excitation occurs inside or outside the ffRF. By complementing the ffRF, the fbRF that we identify here could contribute to predictive processing.

To characterize the ffRF, we mapped receptive field locations of layer 2/3 (L2/3) excitatory neurons in primary visual cortex (V1) of awake, head-fixed mice using two-photon calcium imaging (Fig. 1a). The centre of the ffRF of a given neuron was determined using circular patches of drifting gratings presented individually at different locations (Fig. 1b). To estimate the size of the ffRF, we obtained a size-tuning function by varying the diameter of the grating (Fig. 1c) centred on the ffRF of the neuron (Methods). The responses were maximal for gratings of 13.1 ± 0.4° in diameter and were suppressed with increasing grating size (Fig. 1c), consistent with previous reports[9–12].

To determine the spatial extent of the suppressive regions, we presented a full-field grating in which a portion was masked by a circular grey patch (Fig. 1a, b). We reasoned that the response of the neuron would partially recover upon placing the grey patch on a suppressive region—that is, when part of the suppressive region is not stimulated. We varied the location of the grey patch along the same grid that was used to determine the location of the ffRF. We obtained two separate population-averaged activity maps—the ffRF map and the map for the suppressive regions—and found that the peak of these two maps overlapped (Fig. 1b). Thus, the largest recovery from suppression occurred when the grey patch was located at the centre of the ffRF.

To obtain a finer measure of the response of a neuron to a grey patch, we placed the patch on the centre of the ffRF and varied its diameter. Even the smallest size we tested (5°) evoked a response that was larger than that to the full-field grating. Notably, neuronal responses first increased and then decreased with increasing size of the grey patch (Fig. 1c). These responses were not due to the sharp edges of the stimuli, because similar responses were observed when the edges were blurred (Extended Data Fig. 1). Thus, the size-tuning function of a grey patch on a full-field grating (inverse stimulus) was similar to that of a grating patch on a grey background (classical stimulus).

[1]Department of Physiology, University of California San Francisco, San Francisco, CA, USA. [2]Howard Hughes Medical Institute, University of California San Francisco, San Francisco, CA, USA. ✉e-mail: andreasjakob.keller@ucsf.edu; massimo@ucsf.edu

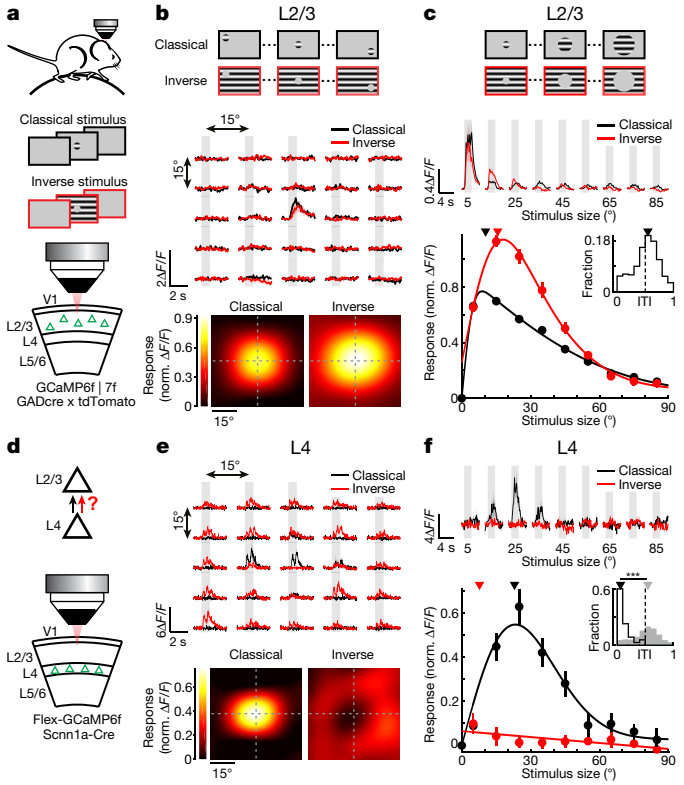


Fig. 1 | Layer-specific responses to inverse stimuli. a, Experimental configuration: Two-photon calcium imaging in excitatory L2/3 neurons (green triangles) of awake head-fixed mice while presenting classical and inverse stimuli. **b**, Top, example trial-averaged responses of an excitatory L2/3 neuron for each stimulus location. Here and in all figures, shaded areas represent stimulus presentation periods. Bottom, population-averaged receptive fields aligned to the centre of the classical ffRF (2,601 excitatory L2/3 neurons in 9 mice). **c**, Top, example trial-averaged responses of a L2/3 neuron for each stimulus diameter. Stimuli are centred on the ffRF. Bottom, population-averaged size-tuning functions, normalized to the maximum response to classical stimuli. In all figures, solid lines show fits to the data and triangles indicate the median preferred size. The inset shows the ITI distribution of L2/3 excitatory neurons, with a median value of 0.54 as indicated by the triangle (the dashed line indicates 0.5; 1,190 excitatory L2/3 neurons in 9 mice). **d**, Schematic of results and experimental configuration. Imaging in excitatory L4 neurons reveals that inverse-tuning in excitatory L2/3 neurons is not simply inherited from L4 neurons. **e**, **f**, Same as **b**, **c**, but for L4 excitatory neurons. The results in **e** are obtained from 24 neurons in 4 mice. In **f**, the ITI for L4 neurons is shown in black (median 0.053; 35 neurons in 6 mice), with the results for L2/3 neurons (**c**) shown in grey. Comparison was performed using a two-sided Wilcoxon rank-sum test; $^{***}P = 1.5 \times 10^{-18}$. In all plots, traces or data points represent the mean and shading or error bars show the s.e.m.

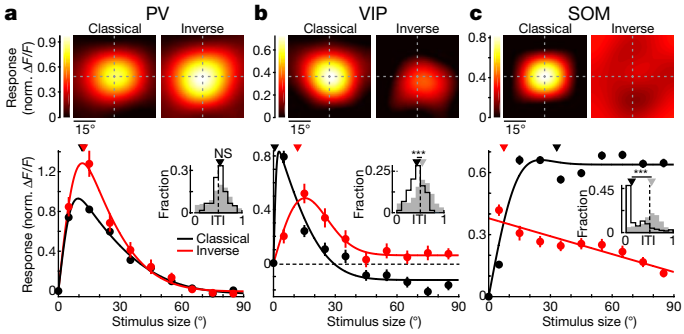


Fig. 2 | Neuron-type specific response to inverse stimuli. a, Top, population-averaged receptive field for PV neurons aligned to the centre of ffRF (82 neurons in 6 mice). Bottom, population-averaged size-tuning functions, normalized to the maximum response to classical stimuli. The inset shows the ITI distribution, with a median of 0.53 as indicated by the triangle (60 neurons in 7 mice). The results for L2/3 neurons are shown in grey (Fig. 1c), and comparison was performed using a two-sided Wilcoxon rank-sum test; NS, $P = 0.79$. **b**, Same as in **a** but for VIP neurons. Top, results from 126 neurons in 4 mice. The median ITI (bottom, inset) is 0.42 (74 neurons in 8 mice). The results for L2/3 neurons are shown in grey (Fig. 1c), and comparison was performed using a two-sided Wilcoxon rank-sum test; $^{***}P = 1.1 \times 10^{-5}$. **c**, Same as in **a** but for SOM neurons. Top, results from 315 neurons in 5 mice. The median ITI (bottom, inset) is 0.12 (179 neurons in 9 mice). The results for L2/3 neurons are shown in grey (Fig. 1c), and comparison was performed using a two-sided Wilcoxon rank-sum test; $^{***}P = 1.1 \times 10^{-42}$.

These results show that L2/3 excitatory neurons are excited by both classical and inverse stimuli centred on their ffRF. Classical and inverse stimuli were mutually antagonistic—the responses to both stimuli together (for example, full-field gratings) were smaller than the responses to either of the stimuli alone (Fig. 1c). We defined a neuron as inverse-tuned if its response to at least one inverse stimulus of any size centred on its ffRF was larger than its response to a full-field stimulus (Methods). Of the visually responsive excitatory neurons in L2/3, 79% were inverse-tuned (943 of 1,190 neurons in 9 mice) (Fig. 1a–c, Extended Data Figs. 2, 3a). We then computed the inverse-tuning index (ITI), in which 0.5 denotes an equal response to both classical and inverse stimuli of the preferred size; 0 indicates a response to classical stimuli only and 1 indicates a response to inverse stimuli only (Methods). The ITI distribution of L2/3 excitatory neurons was unimodal, with a mean of 0.52 ± 0.01 (mean ± s.e.m.) (Fig. 1c).

Surrounding the ffRF of inverse-tuned neurons is a region that is either suppressive or excitatory depending on whether it is stimulated in the presence or in the absence of a stimulus in the ffRF, respectively. We used classical and inverse stimuli to compare the tuning properties of the surrounding excitatory region with those of the ffRF. On average, the orientation tuning to inverse stimuli was sharper than that to classical stimuli (Extended Data Fig. 3b–f). Moreover, individual neurons were not necessarily tuned to the same orientation when stimulated by classical or inverse stimuli (Extended Data Fig. 3g). We determined the interaction between the surrounding region and the ffRF in neurons with a similar orientation preference by independently varying the contrast of simultaneously presented classical and inverse stimuli. At matching contrasts greater than 13%, the interaction between the surrounding region and the ffRF was antagonistic (Extended Data Fig. 3h).

To address whether inverse tuning in L2/3 is inherited from earlier stages of cortical processing, we measured the responses of excitatory neurons in layer 4 (L4) to classical and inverse stimuli (Fig. 1d–f, Extended Data Fig. 3i). In contrast to L2/3 neurons, the suppressive regions of L4 neurons surrounded their ffRF, creating a ring around the centre (Fig. 1e; the absence of a suppressive ring around the ffRF of L2/3 neurons was not due to insufficient spatial resolution of the mapping stimuli; Extended Data Fig. 4). Further, the responses of L4 neurons to inverse stimuli placed on the centre of their ffRF decreased monotonically with stimulus size, which is consistent with the progressive reduction of feedforward drive (Extended Data Fig. 2) and is again different from L2/3 neurons (compare Fig. 1c and Fig. 1f). Overall, the spatial organization of suppressive regions of L4 neurons is consistent with previous models and observations[7,13] and is distinct from that of L2/3 neurons. Inverse tuning in L2/3 excitatory neurons is therefore not inherited from L4 neurons.

Sources of input to L2/3 neurons are specific to the neuron type[11,14]. To address whether inverse tuning is also present in L2/3 inhibitory neurons, we characterized the responses of the three major classes of cortical inhibitory neurons—parvalbumin-expressing (PV), vasoactive-intestinal-peptide-expressing (VIP) and somatostatin-expressing (SOM) neurons—to classical and inverse stimuli (Fig. 2, Extended Data Fig. 3j–l). Both PV and VIP neurons

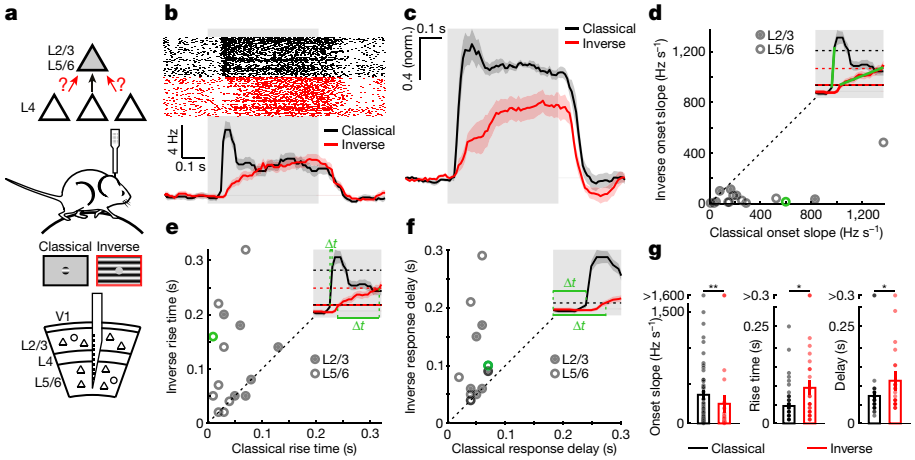


Fig. 3 | Slow and delayed responses to inverse stimuli. a, Schematic of results and experimental configuration for extracellular recordings in V1. **b**, Responses of a L5/6 unit to classical and inverse stimuli centred on its ffRF. Top, raster plot (1,000 trials for each stimulus); bottom, peristimulus time histogram (PSTH; 10-ms bins). **c**, Population-averaged PSTHs normalized to the average response to classical stimuli (15 units in 4 mice). **d**, Onset slope of the response to classical and inverse stimuli. The green symbol represents the example unit from **b**. The inset shows the PSTH from **b** to illustrate the difference in slope between the responses to classical and inverse stimuli. The dotted lines indicate the lower and upper thresholds used to compute the slopes. Comparisons were performed using a two-sided Wilcoxon signed-rank test; $P = 1.8 \times 10^{-4}$. **e**, Same as **d** but for response rise time. Results are for 15 units in 4 mice. Two-sided Wilcoxon signed-rank test; $P = 7.8 \times 10^{-3}$. **f**, Same as **d** but for response delay. Comparisons were performed using a two-sided Wilcoxon signed-rank test; $P = 9.8 \times 10^{-4}$. **g**, Mean onset slopes (left), rise times (middle) and delays (right) of independently tuned units. Classical, 51 units in 8 mice; inverse, 29 units in 8 mice. Two-sided Wilcoxon rank-sum test; onset slope, $^{**}P = 1.1 \times 10^{-3}$; rise time, $^{*}P = 0.017$; delay, $^{*}P = 0.031$.

showed surround suppression to classical stimuli, consistent with previous reports[11,15]. Furthermore, both PV and VIP neurons responded to inverse stimuli centred on their ffRF (Fig. 2a, b) and showed size-tuning functions to inverse stimuli that peaked well above their responses to the largest classical stimuli. By contrast, SOM neurons showed almost no surround suppression to classical stimuli[11,15], poor and spatially diffuse responses to inverse stimuli (Fig. 2c), and none of their responses to inverse stimuli was greater than their response to the largest classical stimuli. Therefore, similar to L2/3 excitatory neurons, most PV and VIP neurons were inverse-tuned whereas SOM neurons were not (Fig. 2).

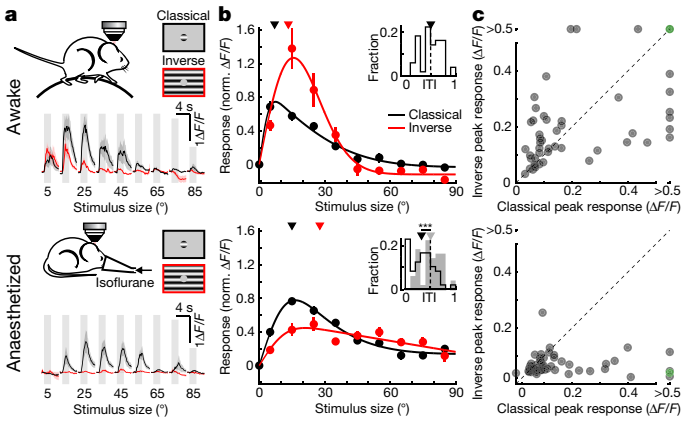


Fig. 4 | Anaesthesia preferentially reduces responses to inverse stimuli. a, Top, responses of an example neuron to classical and inverse stimuli in an awake mouse. Bottom, responses of the same example neuron under isoflurane anaesthesia. **b**, Population-averaged size-tuning functions in awake (top) and anaesthetized (bottom) mice, normalized to the maximum awake response to classical stimuli. The insets show the ITI distributions with median values of 0.50 (top) and 0.32 (bottom). In the bottom inset, black represents neurons from anaesthetized mice and grey from awake mice, compared using a two-sided Wilcoxon signed-rank test; $^{***}P = 1.5 \times 10^{-5}$. The same 49 excitatory L2/3 neurons were measured for both, in 5 mice. **c**, Peak responses of inverse-tuned neurons in awake (top) and anaesthetized (bottom) mice. The green symbol represents the example neuron from **a**. The same 49 excitatory L2/3 neurons in 5 mice were measured as in **b**. Peak responses to classical and inverse stimuli were compared using a two-sided Wilcoxon signed-rank test; top, $P = 0.96$; bottom, $P = 2.0 \times 10^{-5}$.

Although L2/3 neurons do not directly inherit inverse tuning from L4 neurons, the excitatory region surrounding their ffRF could still be generated via feedforward inputs from L4 neurons with spatially offset ffRFs (Fig. 3a). Alternatively, the excitatory region surrounding the ffRFs of L2/3 neurons can be generated via feedback projections[16,17]. We therefore compared the latency of the responses of inverse-tuned neurons to classical and inverse stimuli using extracellular electrophysiological recordings (Fig. 3a) and isolated single units throughout cortical layers, including infragranular layers (L5/6). A large fraction of infragranular units (50%, 60 of 119 units) were also inverse-tuned (Extended Data Fig. 5). In inverse-tuned units recorded in both supra- and infragranular layers, the time course of the responses to classical and inverse stimuli were markedly different (Fig. 3b, c). Whereas the response to classical stimuli showed a fast initial transient followed by a plateau, the response to inverse stimuli slowly progressed towards steady state (Fig. 3c–e) and was delayed relative to the classical response (50 ± 20 ms; mean ± s.e.m.; 15 units) (Fig. 3f). The same biases were observed when comparing the response dynamics to classical stimuli in all responsive units with those to inverse stimuli in inverse-tuned units (Fig. 3g). The difference in latencies and the slower dynamics of responses to inverse stimuli suggest that the excitatory region surrounding the ffRF of L2/3 neurons is unlikely to emerge from the feedforward pathway.

We next determined whether responses to inverse stimuli depend on feedback projections from higher visual areas (HVAs). The effect of anaesthesia on sensory responses has been proposed to be stronger in HVAs than in V1[18]. We therefore compared the effect of isoflurane on the responses to classical stimuli in V1 and in HVAs that had been identified beforehand using wide-field intrinsic imaging (Extended Data Fig. 6a–c). Anaesthesia suppressed responses to a greater extent in HVAs than in V1 (Extended Data Fig. 6d–f). If the responses of L2/3 neurons to inverse stimuli rely on feedback projections, they should be more sensitive to anaesthesia than the responses to classical stimuli. Consistent with this, anaesthesia preferentially suppressed responses to inverse rather than classical stimuli in inverse-tuned neurons (Fig. 4),

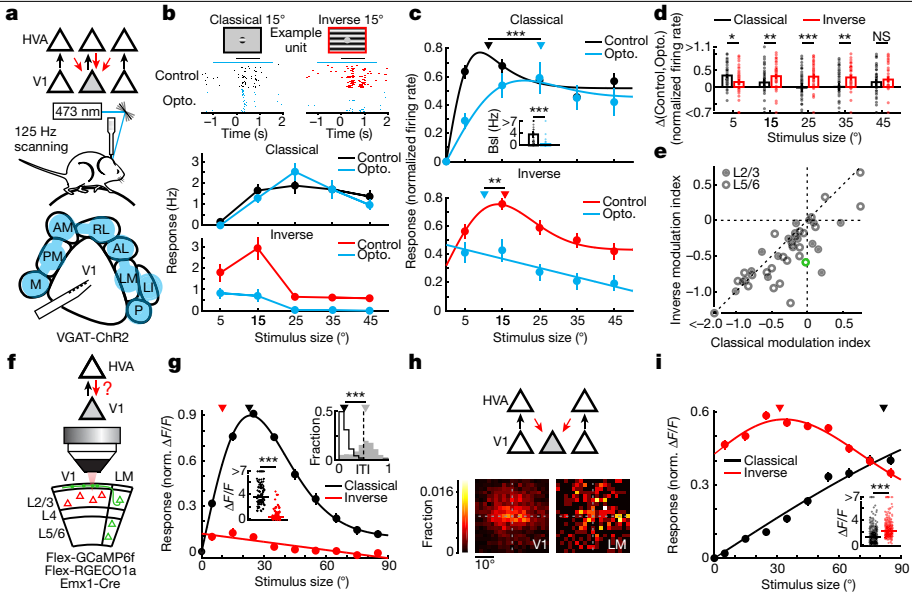


Fig. 5 | Higher visual areas contribute to inverse tuning in V1. a, Schematic of results and experimental configuration for the optogenetic silencing of HVAs. **b**, Top, raster plot of an example L5/6 unit (30 trials each). Black horizontal lines represent the period of stimulus presentation; blue horizontal lines represent period of HVA silencing. 'Control' and 'opto.' indicate trials without and with optogenetic silencing of HVAs, respectively. Bottom, example size-tuning function with or without HVA silencing. **c**, Population-averaged size-tuning function to classical (top) or inverse (bottom) stimuli, with or without HVA silencing, normalized to the maximum control response to classical stimuli. Responses under control and silenced conditions were compared using two-sided Wilcoxon signed-rank tests; classical, $^{***}P = 3.2 \times 10^{-4}$; inverse, $^{**}P = 3.0 \times 10^{-3}$. The inset shows the mean baseline firing rate (bsl) with or without HVA silencing, compared using a two-sided Wilcoxon rank-sum test; $^{***}P = 5.0 \times 10^{-10}$; 44 units in 12 mice. **d**, Difference in firing rates under control conditions and after HVA silencing. Comparisons were performed using two-sided Wilcoxon signed-rank tests. At 5°, $^{*}P = 8.0 \times 10^{-3}$; at 15°, $^{**}P = 1.4 \times 10^{-3}$; at 25°, $^{***}P = 4.0 \times 10^{-5}$; at 35°, $^{**}P = 3.5 \times 10^{-3}$; at 45°, NS, $P = 0.70$; 44 units in 12 mice. **e**, Optogenetic modulation indices (Methods). The green symbol indicates the example unit from **b**. Comparisons were performed using a two-sided Wilcoxon signed-rank test; $P = 7.4 \times 10^{-4}$; 44 units in 12 mice. **f**, Schematic of results and experimental configuration. LM boutons in V1 are not inverse-tuned as revealed by two-photon calcium imaging. **g**, Population-averaged size-tuning function for LM boutons retinotopically aligned to their putative V1 targets. The left inset shows maximum responses, with horizontal lines denoting the median values. Responses to classical and inverse stimuli were compared using two-sided Wilcoxon signed-rank tests; $^{***}P = 6.1 \times 10^{-16}$. The right inset shows the ITI distribution with a median value of 0.11. Black represents results for LM boutons (87 boutons in 5 mice), grey represents the results for L2/3 neurons (Fig. 1c), compared using a two-sided Wilcoxon rank-sum test, $^{***}P = 2.8 \times 10^{-38}$. **h**, Top, LM boutons respond to inverse stimuli that are centred on their putative V1 targets. Bottom, retinotopic spread of the ffRF of V1 neurons and LM boutons (2,352 neurons and 311 boutons in the same 5 mice). **i**, Population-averaged size-tuning functions of LM boutons not retinotopically aligned with their putative V1 targets. The inset shows maximum responses, with horizontal lines denoting the median values. Responses to classical and inverse stimuli were compared using a two-sided Wilcoxon signed-rank test; $^{***}P = 4.8 \times 10^{-36}$; 362 boutons in 5 mice.

suggesting that the response to inverse stimuli may be driven by feedback projections from HVAs.

To directly test the involvement of feedback projections in inverse tuning, we silenced HVAs by scanning them with a laser to optogenetically activate inhibitory neurons while recording extracellular electrophysiological activity in V1 (Fig. 5a–e, Extended Data Fig. 7a–c). In inverse-tuned units, silencing HVAs reduced both spontaneous activity and responses to small-diameter classical stimuli. In surround-suppressed units, silencing HVAs increased responses to large-diameter classical stimuli (Fig. 5c, d, Extended Data Fig. 8), as has previously been shown[19–21]. The response to inverse stimuli, however, was strongly suppressed (Fig. 5c–e). Upon silencing HVAs, inverse stimuli evoked responses that decreased with increasing size of the grey patch, reminiscent of the responses of L4 neurons. These effects could not be explained by a direct effect of scattered laser light on V1 (Extended Data Fig. 7d–g), nor by the activation of putative inhibitory neurons in the HVA with long-range axonal projections targeting V1, as these projections—consistent with the findings of a previous study[22]—were rare (Extended Data Fig. 7h–l).

To test whether distinct HVAs contribute equally to inverse responses in V1, we silenced individual HVAs while recording single-unit responses in V1 to classical and inverse stimuli (Extended Data Fig. 9). Although the silencing of several visual areas reduced the response to inverse stimuli, the strongest stimulus-specific effect on inverse responses occurred when silencing the lateromedial visual area (LM).

To address whether inverse tuning is directly inherited from inverse-tuned neurons in the LM, we determined the response properties of LM axonal boutons in layer 1 of V1 while mapping the retinotopic coordinates of the V1 site (Fig. 5f–i, Extended Data Fig. 10a–h). LM boutons for which the receptive fields were centred on the retinotopic coordinates of the imaged V1 site showed surround suppression to classical stimuli and were not inverse-tuned (Fig. 5g). When presenting inverse stimuli centred on the receptive field of the LM boutons, the response decreased with increasing diameter of the grey patch, as seen in L4 neurons (Fig. 1f). This was not a general property of LM neurons, because directly imaging cell bodies in the LM showed inverse tuning in some neurons and in the population average (Extended Data Fig. 10i–k). We therefore established that V1 neurons do not directly inherit inverse tuning from the LM.

Inverse tuning of L2/3 neurons could result from LM inputs that, although not inverse-tuned, have spatially offset receptive fields relative to those of L2/3 neurons. These LM inputs would respond to an inverse stimulus centred on the V1 retinotopic coordinates because their receptive field, being offset relative to the grey patch, would be stimulated by the grating. We mapped the spatial offset of the receptive field of LM boutons relative to the retinotopic coordinates of the V1 sites. The centres of the receptive fields of LM boutons showed a wide scatter relative to the retinotopic coordinates of the V1 site, larger than the scatter of ffRF centres of L2/3 neurons at the V1 site (Fig. 5h, Extended Data Fig. 10f, g), consistent with the findings of a previous

study[23]. LM boutons with spatially offset receptive-field centres therefore converge on a given retinotopic site in V1. If these LM inputs contribute to the inverse response of L2/3 neurons, they should respond to inverse stimuli centred on the V1 site. This was indeed the case on average, and a large fraction of these boutons significantly responded to inverse stimuli with a preference for small sizes (Fig. 5i, Extended Data Fig. 10h); this is consistent with the inverse size-tuning function in L2/3 neurons (Fig. 1c). In addition, the response of these boutons to classical stimuli of progressively larger diameter centred on the V1 site increased gradually (Fig. 5i), consistent with their receptive fields being offset relative to the centre of the stimulus. Inverse tuning in L2/3 V1 neurons is therefore likely to result from the feedback of non-inverse-tuned neurons in HVAs, which have receptive fields that are offset relative to the ffRF centres of the V1 neurons on which they converge.

Our results demonstrate that feedback projections to V1 neurons generate a second, distinct excitatory receptive field that surrounds the ffRF. This feedback receptive field (fbRF) is absent in L4 and emerges along the laminar processing hierarchy in the supra- and infragranular layers of V1. The fbRF and the ffRF are mutually antagonistic, such that neurons respond when a stimulus is presented in either the fbRF or the ffRF but not in both together, effectively performing an exclusive-OR operation. The suppression of responses to stimuli in the ffRF by surrounding stimuli is a well-established phenomenon that enables neurons to report differences in stimulus features between the excited region inside the ffRF and its surround[7,8,11,12,24–27]. Neurons with an excitatory fbRF report differences in stimulus features regardless of whether the excited region is located inside or outside the ffRF. We propose that SOM inhibitory neurons, which—in contrast to PV and VIP neurons—respond poorly to inverse stimuli while responding robustly to large stimuli covering both fbRFs and ffRFs, could mediate the mutual antagonism, consistent with their role in surround suppression[11].

In addition to HVAs, local excitation within V1 may also contribute to the generation of the fbRF[28,29]. In any case, the fbRF may underlie phenomena such as filling-in or illusory contours in which the stimulus in the ffRF is absent, weak or obstructed[3,5,30–32] and may account for contextual modulation[8,12,24,27,33], detection of borders[6,7] or pop-out effects[34].

The antagonism between ffRF and fbRF is reminiscent of models of predictive processing[35–37] in which bottom-up information about the stimulus is compared with top-down predictions, such that only differences between prediction and stimulus identity are represented. Surround suppression has been interpreted within this framework[36]. If the visual stimulus that surrounds the ffRF provides a correct estimate of the stimulus in the ffRF, the response can be suppressed as there is no difference between stimulus prediction and stimulus identity. With inverse tuning, the framework of predictive processing generalizes to stimuli within and outside of the ffRF owing to the presence of a fbRF. Independent of any conceptual framework, the presence of a fbRF generated by feedback projections probably accounts for several aspects of sensory processing along the cortical hierarchy.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2319-4.

1. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160**, 106–154 (1962).
2. Pennartz, C. M. A., Dora, S., Muckli, L. & Lorteije, J. A. M. Towards a unified view on pathways and functions of neural recurrent processing. *Trends Neurosci.* **42**, 589–603 (2019).
3. von der Heydt, R., Peterhans, E. & Baumgartner, G. Illusory contours and cortical neuron responses. *Science* **224**, 1260–1262 (1984).
4. Shen, Z.-M., Xu, W.-F. & Li, C.-Y. Cue-invariant detection of centre-surround discontinuity by V1 neurons in awake macaque monkey. *J. Physiol. (Lond.)* **583**, 581–592 (2007).
5. Fiorani Júnior, M., Rosa, M. G. P., Gattass, R. & Rocha-Miranda, C. E. Dynamic surrounds of receptive fields in primate striate cortex: a physiological basis for perceptual completion? *Proc. Natl Acad. Sci. USA* **89**, 8547–8551 (1992).
6. Rossi, A. F., Desimone, R. & Ungerleider, L. G. Contextual modulation in primary visual cortex of macaques. *J. Neurosci.* **21**, 1698–1709 (2001).
7. Jones, H. E., Grieve, K. L., Wang, W. & Sillito, A. M. Surround suppression in primate V1. *J. Neurophysiol.* **86**, 2011–2028 (2001).
8. Schnabel, U. H. et al. Figure-ground perception in the awake mouse and neuronal activity elicited by figure-ground stimuli in primary visual cortex. *Sci. Rep.* **8**, 17800 (2018).
9. Hübener, M. Mouse visual cortex. *Curr. Opin. Neurobiol.* **13**, 413–420 (2003).
10. Niell, C. M. & Stryker, M. P. Highly selective receptive fields in mouse visual cortex. *J. Neurosci.* **28**, 7520–7536 (2008).
11. Adesnik, H., Bruns, W., Taniguchi, H., Huang, Z. J. & Scanziani, M. A neural circuit for spatial summation in visual cortex. *Nature* **490**, 226–231 (2012).
12. Angelucci, A. et al. Circuits and mechanisms for surround modulation in visual cortex. *Annu. Rev. Neurosci.* **40**, 425–451 (2017).
13. Walker, G. A., Ohzawa, I. & Freeman, R. D. Asymmetric suppression outside the classical receptive field of the visual cortex. *J. Neurosci.* **19**, 10536–10553 (1999).
14. Callaway, E. M. Cell type specificity of local cortical connections. *J. Neurocytol.* **31**, 231–237 (2002).
15. Dipoppa, M. et al. Vision and locomotion shape the interactions between neuron types in mouse visual cortex. *Neuron* **98**, 602–615.e8 (2018).
16. Mignard, M. & Malpeli, J. G. Paths of information flow through visual cortex. *Science* **251**, 1249–1251 (1991).
17. Wang, Q., Sporns, O. & Burkhalter, A. Network analysis of corticocortical connections reveals ventral and dorsal processing streams in mouse visual cortex. *J. Neurosci.* **32**, 4386–4399 (2012).
18. Lamme, V. A. F., Zipser, K. & Spekreijse, H. Figure-ground activity in primary visual cortex is suppressed by anesthesia. *Proc. Natl Acad. Sci. USA* **95**, 3263–3268 (1998).
19. Nassi, J. J., Lomber, S. G. & Born, R. T. Corticocortical feedback contributes to surround suppression in V1 of the alert primate. *J. Neurosci.* **33**, 8504–8517 (2013).
20. Nurminen, L., Merlin, S., Bijanzadeh, M., Federer, F. & Angelucci, A. Top-down feedback controls spatial summation and response amplitude in primate visual cortex. *Nat. Commun.* **9**, 2281 (2018).
21. Vangeneugden, J. et al. Activity in lateral visual areas contributes to surround suppression in awake mouse V1. *Curr. Biol.* **29**, 4268–4275.e7 (2019).
22. McDonald, C. T. & Burkhalter, A. Organization of long-range inhibitory connections with rat visual cortex. *J. Neurosci.* **13**, 768–781 (1993).
23. Marques, T., Nguyen, J., Fioreze, G. & Petreanu, L. The functional organization of cortical feedback inputs to primary visual cortex. *Nat. Neurosci.* **21**, 757–764 (2018).
24. Kapadia, M. K., Westheimer, G. & Gilbert, C. D. Spatial distribution of contextual interactions in primary visual cortex and in visual perception. *J. Neurophysiol.* **84**, 2048–2062 (2000).
25. Maffei, L. & Fiorentini, A. The unresponsive regions of visual cortical receptive fields. *Vision Res.* **16**, 1131–1139 (1976).
26. Self, M. W. et al. Orientation-tuned surround suppression in mouse visual cortex. *J. Neurosci.* **34**, 9290–9304 (2014).
27. Keller, A. J. et al. A disinhibitory circuit for contextual modulation in primary visual cortex. Preprint at https://www.biorxiv.org/content/10.1101/2020.01.31.929166v2 (2020).
28. Gilbert, C. D. & Wiesel, T. N. Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J. Neurosci.* **9**, 2432–2442 (1989).
29. Fitzpatrick, D. Seeing beyond the receptive field in primary visual cortex. *Curr. Opin. Neurobiol.* **10**, 438–443 (2000).
30. Peterhans, E. & Von der Heydt, R. Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *J. Neurosci.* **9**, 1749–1763 (1989).
31. Smith, F. W. & Muckli, L. Nonstimulated early visual areas carry information about surrounding context. *Proc. Natl Acad. Sci. USA* **107**, 20099–20103 (2010).
32. Grosof, D. H., Shapley, R. M. & Hawken, M. J. Macaque V1 neurons can signal 'illusory' contours. *Nature* **365**, 550–552 (1993).
33. Seriès, P., Lorenceau, J. & Frégnac, Y. The "silent" surround of V1 receptive fields: theory and experiments. *J. Physiol. Paris* **97**, 453–474 (2003).
34. Knierim, J. J. & van Essen, D. C. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophysiol.* **67**, 961–980 (1992).
35. Bastos, A. M. et al. Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711 (2012).
36. Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
37. Keller, G. B. & Mrsic-Flogel, T. D. Predictive processing: a canonical cortical computation. *Neuron* **100**, 424–435 (2018).

Methods

Mice

All experimental procedures were approved by the regulation of the Institutional Animal Care and Use Committee (IACUC, AN179056) of the University of California, San Francisco. Mice of either sex were kept on a C57BL/6 background (except VIP-IRES-Cre) and were of the following genotype:

Gad2-IRES-cre (GAD2$^{tm2(cre)Zjh}$; JAX 010802) × Ai14 (Gt(ROSA)26Sor$^{tm14(CAG-tdTomato)Hze}$; JAX 007914) for imaging of L2/3 excitatory neurons (9 mice; Figs. 1a–c, 4, Extended Data Figs. 1, 3, 4); Emx1-IRES-cre (Emx1$^{tm1(cre)Krj}$; JAX 005628) for imaging L2/3 excitatory neurons and axons from the LM (5 mice; Fig. 5f–i, Extended Data Fig. 10a–h); Gad2-IRES-cre (GAD2$^{tm2(cre)Zjh}$; JAX 010802) for imaging L2/3 neurons and labelling inhibitory projections (8 mice; Extended Data Figs. 6, 7h–k, 10i–k); PV-cre (Pvalb$^{tm1(cre)Arbr}$; JAX 017320) × Ai14 (Gt(ROSA)26Sor$^{tm14(CAG-tdTomato)Hze}$; JAX 007914) for imaging of L2/3 PV neurons (7 mice; Fig. 2a); VIP-IRES-cre (Vip$^{tm1(cre)Zjh}$; JAX 010908) × Ai14 (Gt(ROSA)26Sor$^{tm14(CAG-tdTomato)Hze}$; JAX 007914) for imaging of L2/3 VIP neurons (8 mice; Fig. 2b); Sst-IRES-cre (Sst$^{tm2.1(cre)Zjh}$; JAX 028864) × Ai14 (Gt(ROSA)26Sor$^{tm14(CAG-tdTomato)Hze}$; JAX 007914) for imaging of L2/3 SOM neurons (5 mice; Fig. 2c); Scnn1a-Tg3-cre (Tg(Scnn1a-cre)3Aibs/J; JAX 009613) and Scnn1a-Tg3-cre (Tg(Scnn1a-cre)3Aibs/J; JAX 009613) × Ai148 (Igs7$^{tm148.1(tetO-GCaMP6f,CAG-tTA2)Hze}$; JAX 030328) for imaging L4 excitatory neurons (5 mice and 1 mouse, respectively; Fig. 1d–f, Extended Data Fig. 4d) and VGAT-ChR2-EYFP (Tg(Slc32a1-COP4*H134R/EYFP)8Gfng/J; JAX 014548) for electrophysiology and optogenetic inhibition experiments (20 mice; Figs. 3, 5a–e, Extended Data Figs. 5, 7a–g, 8, 9). The mice were housed on a reverse light cycle (light/dark cycle 12/12 h). At the start of the experiments, all mice were between 2 and 9 months old.

Viruses

We injected the following viruses: AAV2/1.ef1a.GCaMP6f.WPRE (FMI Vector Core Facility), AAV2/1.ef1a.DIO.GCaMP6f.WPRE (FMI Vector Core Facility), AAV2/1.CAG.CGaMP6f (Janelia Vector Core), AAV2/9.syn.GCaMP7f (Addgene), AAV1.Syn.Flex.NES-jRGECO1a.WPRE.SV40 (Addgene) and AAVretro.CAG.Flex.tdTomato (Addgene). Viruses were diluted to use titres of approximately 5×10^{12} genome copies per ml and 50 nl was injected at each injection site (3 to 5 sites for two-photon experiments and 1 site for anatomy experiments) and each depth (2 from 350 to 200 µm below the pial surface for two-photon calcium imaging experiments; 4 from 650 to 200 µm below the pial surface for the anatomy experiments and two-photon recordings of LM boutons).

Surgery

Mice were anaesthetized with 2% isoflurane or with a mixture of fentanyl (West-Ward Pharmaceuticals, 0.05 mg kg^{-1}), midazolam (Akorn, 5.0 mg kg^{-1}) and dexmedetomidine (Zoetis, 0.5 mg kg^{-1}), injected subcutaneously. The body temperature of the mice was monitored and kept constant. To prevent the eyes from drying, a layer of lubricant ointment (Rugby) was applied. The skin above the skull was disinfected with povidone iodine. For mice prepared for intrinsic optical imaging (those needed for two-photon calcium imaging in HVAs or in LM boutons, and for all electrophysiology experiments), the bone over the right visual cortex was thinned, the exposed skull was covered with a thin layer of glue (Krazy Glue) and a headplate was attached using dental cement (Ortho-Jet Powder, Lang). The mice were then allowed to recover for several days before any other surgical or experimental procedures. For two-photon experiments, a craniotomy was made over the right visual cortex (3–4.5 mm in diameter) and viruses were injected with a micropump (UMP-3, World Precision Instruments) at a rate of 2 nl s^{-1}. The craniotomy was then sealed with a glass coverslip using cyanoacrylate glue and, if not already present, a headplate was attached. For electrophysiology experiments, a small craniotomy was performed (approximately 0.3 mm in diameter) guided by the activity maps of the visual cortex obtained by intrinsic optical imaging. After the recording, the mouse was either perfused for histology or its skull was protected with Kwik-Cast (World Precision Instruments) for the next experiment. For anatomical experiments, the skin was sutured after the viral injection using 6-0 suture silk (Fisher Scientific NC9134710). To reverse the anaesthesia induced by the fentanyl/midazolam/dexmedetomidine mixture, a mixture of naloxone (Hospira, 1.2 mg kg^{-1}), flumazenil (West-Ward Pharmaceuticals, 0.5 mg kg^{-1}) and atipamezol (Zoetis, 2.5 mg kg^{-1}) was injected subcutaneously after the surgical procedures.

Visual stimulation

Visual stimuli were generated using the open-source Psychophysics Toolbox[38] based on MATLAB (MathWorks). Stimuli were presented at a distance of 15 cm to the left eye on a gamma-corrected LED-backlit LCD monitor (Dell) with a mean luminance of 20 cd m^{-2}. For two-photon experiments using a resonant scanner, the power source of the LED backlight of the monitor was synchronized to the resonant scanner turnaround points (when data were not acquired) to minimize light leak from the monitor[39]. We presented drifting sinusoidal gratings (2 Hz, 0.04 cycles per degree, 100% contrast) unless stated otherwise. The trial structure of all stimulus sessions (for example, for receptive field mapping, size-tuning experiments and so on) was block randomized (the block size was given by the total number of parameter combinations). In all raster plots (Figs. 3, 5, Extended Data Fig. 7), we separated stimulus conditions for clarity.

Intrinsic imaging. To estimate the visual area locations and their retinotopic maps using intrinsic imaging, we presented a narrow white bar (5°) on a black background, slowly drifting (10° per second) in one of the cardinal directions (10 to 20 trials per direction). In addition, we presented 25° patches of gratings at different retinotopic locations (usually one nasal and one temporal, 20 trials each). Gratings were presented for 2 s at 8 different directions (0.25 s each) followed by 13 s of grey screen.

Receptive field mapping. Stimuli consisted of either a 20° circular grating patch on a grey screen (classical stimulus) or a 20° grey circular patch on a full-field grating (that is, large gratings covering the entire screen, approximately 120 × 90°; inverse stimulus) with a 15° spacing between the centre of the patches (regular grid). For two-photon calcium-imaging experiments, stimuli were presented for 1 s at a single direction or for 2 s at the four cardinal directions (0.5 s each). Stimulation periods were interleaved by 2 s of grey screen. We recorded 5 to 10 trials per stimulus condition. For electrophysiological experiments, stimuli were presented for 0.5 s at a single direction interleaved by 1 s of grey screen. We recorded 20 trials per stimulus condition. In addition, we used a finer grid of grating patches in a subset of experiments (patches of 10° with a spacing of 5°; Extended Data Fig. 4a, b).

Orientation tuning. We presented gratings of at least 15° diameter drifting in 8 directions (5 to 10 trials and 20 trials per direction for two-photon calcium-imaging and electrophysiology experiments, respectively). For the experiments shown in Extended Data Fig. 3, we additionally presented inverse gratings drifting in 8 directions, centred on the classical ffRF. The stimulus presentation time was 1 s, interleaved with 1.5 to 2 s of grey screen.

Size tuning. Patches of gratings and inverse gratings were displayed at 9 different sizes, equally spaced from 5 to 85° in diameter (10 trials per size; for two-photon experiments) or at 5 different sizes, equally spaced from 5 to 45° in diameter (20 to 30 trials per size; for electrophysiology experiments), centred on the ffRF. Stimulation time was either 2 s interleaved by 4 s of grey screen (for two-photon experiments) or 1 s interleaved by 1.5 s of grey screen (for electrophysiology experiments).

Trials with optogenetic stimulation had an additional 1 s pre-stimulus and 0.5 s post-stimulus grey screen during which the optogenetic light source was turned on and the total number of trials was doubled (see 'Optogenetics'). In addition, we blurred the edge of the patches using a sigmoid function increasing from 1% to 99% over 10° in a subset of experiments (Extended Data Fig. 1). All other parameters were the same as for the size tuning described above.

Contrast tuning. We simultaneously presented classical and inverse stimuli with several test contrasts (0, 2^{-6}, 2^{-5}, …, 1). Stimuli were presented for 2 s interleaved by 4 s of grey screen (10 trials per stimulus combination).

Response dynamics. To estimate the temporal response profile to inverse stimuli (Fig. 3), we presented patches of gratings and inverse gratings at a single size (1,000 trials each). These gratings were presented either at 15° or 20°, for 0.5 s interleaved by 1 s of grey screen. The initial phase of the drifting gratings was randomized to avoid overestimating the onset delay of the response for simple-cell-like receptive fields.

Behavioural monitoring

All mice were habituated (3 to 5 days) to the experimental setups before starting experiments. During all awake experiments, we recorded the positions of the left eye using a CMOS camera (DMK23UM021, Imaging Source) with a 50-mm lens (M5018-MP, Moritex), tracked the running speed of the mouse, and monitored its general behaviour using a webcam (LifeCam Cinema 720p HD, Microsoft). Excluding eye-movement or running trials did not affect the results. For experiments under anaesthesia that followed awake experiments (Fig. 4, Extended Data Fig. 6), mice were anaesthetized with isoflurane (approximately 1% in O_2) delivered with a nose cone. After induction of anaesthesia, the body temperature of the mice was monitored and kept constant. To ensure an adequate depth of anaesthesia, we tracked the left eye and monitored general behaviour.

Intrinsic optical imaging

We used intrinsic optical imaging to identify the centre of the V1 or the locations of HVAs. We sedated the mice with chlorprothixene (0.7 mg kg^{-1}) then lightly anaesthetized with isoflurane (0.5 to 1% in O_2) delivered through a nose cone. The rectal temperature was monitored and maintained at 37 °C. We illuminated the visual cortex with 625-nm light from two LED light sources (M625F2, Thorlabs) using 1.5-mm light fibres (FP1500URT, Thorlabs). The intrinsic optical signal was measured with an Olympus MVX stereo-macroscope using a narrow bandpass filter (700/13 nm BrightLine, Semrock). We acquired the images at 10 Hz with a CCD camera (Orca-Flash 4.0 v2, Hamamatsu) using custom-written software in LabVIEW (National Instruments).

Two-photon calcium imaging

Imaging was performed using either a galvanometric-scanner-based movable objective microscope (MOM) (Sutter) or a resonant-scanner-based (8 kHz) Bergamo II two-photon microscope (Thorlabs), both controlled by ScanImage (Vidrio). Using the MOM system, we acquired images of 128 × 128 pixels at a single depth at a frame rate of 5.92 Hz. With the Bergamo II microscope, we acquired images of 380 × 512 pixels at 1 or 4 depths at frame rates of 40 Hz or 8 Hz, respectively. We obtained similar results with both systems, so all data were pooled. The illumination light source was a Ti:sapphire laser (Chameleon Ultra II, Coherent) used at excitation wavelengths of 910 nm for green indicator imaging and 1,040 nm for red indicator imaging. The laser power under the objective (16× , Nikon) never exceeded 50 mW (laser pulse width 140 fs at a repetition rate of 80 MHz).

Electrophysiology

We performed extracellular recordings using multi-electrode silicon probes (A1x32-Edge-5 mm-20-177-A32, NeuroNexus) with 32 channels spaced by 20 µm. The recording electrodes were controlled with micromanipulators (Luigs & Neumann) and coated with DiO lipophilic dyes (Life Technologies) for post-hoc identification of the electrode track. We recorded the bandpass-filtered (0.1–7.5 kHz) signals at 30 kHz using an Intan system (RHD2000 USB Interface Board, Intan Technologies).

Optogenetics

We used the VGAT-ChR2-EYFP mouse line to ensure a homogeneous expression of the opsin. To silence parts of the visual cortex, we used a 473-nm laser (LuxX 473-80, Omicron-Laserage). The light was first guided through a pinhole to collimate the beam, then sent through a long-range focal lens (AC254-300-A, Thorlabs) to focus the light onto the cortical surface (theoretical spot size ≤ 200 µm), before it entered a 2D-galvo system (GVS202, Thorlabs) to direct the light to the regions of interest. The scanners were controlled by custom-written software in LabVIEW and guided by a CMOS camera (DMK23UM021, Imaging Source) with a 50-mm lens (M5018-MP, Moritex). For Fig. 5a–e and Extended Data Figs. 7–9, we defined 8 HVAs (P, LI, LM, AL, RL, AM, PM, and M) on the basis of the intrinsic optical imaging maps established before the optogenetic experiment (see 'Intrinsic optical imaging maps'). These areas were consecutively scanned in a circular manner with a dwell time of ≤ 1 ms per area (resulting in a frequency of 125 Hz for the whole cycle). The laser was briefly shut off each time the beam moved from areas M and P to avoid silencing parts of V1 (see 'Visual stimulation' for timing within a trial). For assessing the role of single HVAs in the generation of inverse tuning, we targeted each area individually (Extended Data Fig. 9). To verify the effectiveness of the silencing using this approach, we performed control recordings by scanning over the recording site in V1 (Extended Data Fig. 7a–c). To measure the spatial extent of silencing, we parked the laser at 5 locations at and around the recording site (800 µm and 400 µm lateral and medial of the recording site and on the recording site itself, randomizing which location to silence for each trial), targeted individual HVAs, or scanned over these 8 HVAs (Extended Data Fig. 7d–g). For experiments scanning over all 8 HVAs, the laser power was set to approximately 0.75 mW mm^{-2} (total power at the surface of the cortex was 3 mW distributed over approximately 4 mm^2 of illuminated HVAs). For experiments targeting individual locations or HVAs, the laser power was set to approximately 4 mW mm^{-2} (the total power at the surface of the cortex was 2 mW).

Histology

Mice were deeply anaesthetized with 5% isoflurane and urethane, and transcardially perfused with PBS followed by 4% paraformaldehyde in PBS. The brain was then embedded in 2–3% agar and 100-µm-thick sections were cut using a microtome (Leica VT1000 S vibratome). Slices were mounted using a Vectashield HardSet mounting medium containing DAPI (H-1500-10, Vector Laboratories H1500). Images were acquired with an Olympus MVX10 MacroView microscope or a Nikon Ti CSU-W1 inverted spinning disk confocal microscope, and analysed using Fiji[40].

For electrophysiology experiments, the penetration depth was estimated post hoc using the DiO track of the electrode (see 'Electrophysiology') and the L4/L5 border was defined on the basis of the DAPI staining. This enabled us to determine which pins of the electrode were located in L5/6. For scatter plots, inverse-tuned units (see 'Data analysis') were defined as L2/3 or L5/6 units if they were above or below this border, respectively.

To identify and quantify inhibitory long-range projections from HVAs to V1 (Extended Data Fig. 7h–l), we injected an AAVretro.CAG.Flex.tdTomato in V1 of GADcre mice. The mice were euthanized approximately three weeks later. The borders between V1 and the HVAs were defined on the basis of the DAPI staining using the thickness of L4. On the basis of these borders and a mouse atlas[41], we defined the location and identity of HVAs. To quantify the number of inhibitory neurons in HVAs projecting to V1, we counted the tdTomato-positive cell bodies in the coronal

slice that contained the centre of the area. This underestimates the difference in the number of projection neurons between V1 and HVAs.

Data analysis

All data were analysed using custom-written code in MATLAB.

Two-photon calcium imaging. We analysed two-photon calcium imaging data as described previously[42]. In brief, data were full-frame registered using custom-written software (https://sourceforge.net/projects/iris-scanning/). We selected the neurons semi-manually, on the basis of mean and maximum projection images. We calculated the raw fluorescence traces as the average fluorescence of all pixels within a selected region of interest for each frame. Fluorescence changes ($\Delta F/F$) were calculated as described elsewhere[43]. All stimulus evoked responses were baseline-subtracted (1 s pre-stimulus interval).

Extracellular recordings. We determined single-unit firing using Kilo-Sort and Phy (https://github.com/cortex-lab/KiloSort). We determined the spike times with 1-ms resolution. Inhibitory units were defined as units for which the firing rate significantly increased ($P < 0.05$) during optogenetic stimulation in the absence of a visual stimulus—that is, during the pre-visual-stimulus baseline. All stimulus-evoked responses were baseline-subtracted (0.5-s pre-stimulus interval).

Response amplitude. The response amplitude to a stimulus was computed as the average response over the duration of the stimulus presentation (excluding the first 0.5 s of each trial for two-photon experiments owing to the delay and slow rise of calcium indicators). Responses were normalized by the maximum response over the relevant stimulus parameter space and then averaged over neurons or units. We defined significant responses as responses that exceeded a z-score of 3.29 (corresponding to $P < 10^{-3}$) or 5.33 (corresponding to $P < 10^{-7}$; for two-photon experiments in L4).

Receptive field mapping. To estimate the centre of the receptive field, we fitted the responses to patches of gratings with a two-dimensional Gaussian. We excluded neurons if they did not have at least one significant trial-averaged response within 10° of their estimated centres (or the closest data point if no stimulus was located within 10°). For the comparison of the average receptive field maps to classical and inverse stimuli (Figs. 1, 2, Extended Data Figs. 4, 5), we included only neurons with at least one significant average response to a classical and an inverse stimulus at any location. To compare regular and fine receptive field mapping (Extended Data Fig. 4), neurons were included only if they responded to both fine and regular grid stimuli and if their estimated receptive field centre (of the regular grid) was within the surface covered by the fine mapping stimuli (see smaller dashed rectangle in Extended Data Fig. 4a). To illustrate the average receptive fields (heat maps in Figs. 1, 2, Extended Data Figs. 4, 5), we used a spline interpolation and smoothed the overall average with a two-dimensional Gaussian filter (10°). We excluded neurons from further analysis (for example, size tuning) if the estimated centres of their ffRFs were not within 10° of the centres of the stimuli presented to establish size tuning, orientation tuning and related properties.

Size tuning. We fitted the data to an integral over a difference of Gaussians. This fit was used to estimate the sizes of the ffRF and fbRF of the neurons. We approximated the size of the ffRF by the size of the patch of gratings evoking the largest response (size-tuning fits were bound to the interval 0.1–90.1°). We excluded neurons from further analysis if they did not respond to at least one classical stimulus of any size. To compare size tuning with sharp and blurred edges, neurons had to respond to at least one classical stimulus of any size for both stimulus types (sharp and blurred) (Extended Data Fig. 1). Surround-suppressed neurons were defined as neurons in which the response to a classical stimulus of any size was significantly larger than that to the largest classical stimulus tested (Extended Data Fig. 5). We calculated the suppression index as the average response over the two largest stimuli presented divided by the maximum response (Extended Data Fig. 8c). The same sizes were used to calculate the suppression index during HVA silencing.

Defining inverse-tuned neurons. Neurons were defined as inverse-tuned if they significantly responded to at least one classical and one inverse stimulus and if their response to at least one inverse stimulus of any size centred on their ffRF was significantly larger than that to a full-field stimulus (or approximated by the response to the largest classical or smallest inverse stimulus presented).

Inverse-tuning index. We defined the ITI as:

$$\mathrm{ITI} = \frac{R_{\mathrm{inv}} - R_{\mathrm{cla}}}{2 \times ((R_{\mathrm{inv}} - R_{\mathrm{ff}}) + (R_{\mathrm{cla}} - R_{\mathrm{ff}}))} + 0.5$$

in which R_{inv} is the maximum response to inverse stimuli, R_{cla} is the maximum response to classical stimuli and R_{ff} is the response to a full-field stimulus.

Orientation tuning. We fitted a circular sum of Gaussians with a peak offset of 180° and equal tuning width (full width at half maximum of the Gaussian fit). We calculated orientation selectivity index (OSI) and direction selectivity index (DSI) as described elsewhere[10]. Classical and inverse stimuli were presented at a fixed stimulus diameter (10°, 15° or 20°). Neurons were excluded from this analysis (Extended Data Fig. 3b–g) if their classical and inverse preferred sizes were not within 10° of the presented stimulus size.

Contrast tuning. Classical and inverse stimuli were presented at a fixed stimulus diameter (10°, 15° or 20°) and at one orientation. Neurons were excluded from this analysis (Extended Data Fig. 3h) if their classical and inverse preferred sizes were not within 10° of the presented stimulus size. Moreover, we excluded neurons if their OSIs were ≥0.3 and if their orientation preference was not within 45° of the presented stimulus orientation. That is, we excluded neurons that were strongly orientation-tuned to the orthogonal orientation.

Response dynamics. To estimate the response delay, rise time and onset slope for classical and inverse stimuli, we binned the spike times in bins of 10 ms and then median-filtered (50 ms) the average traces. We defined the response delay as the first data point after stimulus onset that crossed a z-score threshold of 5.33 (corresponding to $P < 10^{-7}$). Further, we defined the rise time as the interval between the response onset (as estimated for the response delay) and the first time point crossing 75% of the maximum response during stimulus presentation (changing this arbitrary value to 50% or 100% did not affect the results). Finally, we estimated the response onset slope as the fitted slope to the response during the initial rise time. We excluded units for which the responses did not exceed the response threshold defined above. Furthermore, for the population responding to the classical stimulus (Fig. 3g), units were excluded if their preferred classical size was larger than the presented stimulus size (±10°). For the inverse-tuned population (Fig. 3g), units were excluded if their preferred inverse size was smaller than the presented size (±10°). For the inverse-tuned subpopulation of units responding to both (Fig. 3c–f), both classical and inverse sizes were required to be within 10° of the presented stimulus size.

Awake and anaesthetized conditions. Neurons were included in this analysis on the basis of their awake responses (Fig. 4, Extended Data Fig. 6). However, to ensure that the stimuli were also centred on the receptive fields under anaesthesia, neurons were excluded if the

estimated centres of their ffRFs under anaesthesia were not within 10° of the centres of the anaesthetized size-tuning stimuli presented. To estimate the peak response of a neuron under anaesthesia, we used the same size as in the awake condition (±10°).

Size tuning of (non-centred) LM boutons to stimuli centred on their putative V1 targets. Size-tuning stimuli were presented at a location such that the population-averaged centre of the V1 receptive fields was within 10° (Extended Data Fig. 10f, left). LM boutons were excluded from this analysis if they did not respond to any size-tuning stimulus of any size (classical or inverse) and if their estimated centres of their ffRFs were within 10° of the centres of the presented size stimuli (Extended Data Fig. 10h). Hence, only putative offset boutons were included. Additionally, for Fig. 5i, boutons needed to respond to an inverse stimulus of any size (stimulus was not centred on the boutons' receptive fields).

Intrinsic optical imaging maps. We calculated the temporal phase of the Fourier component at the frequency of the bar presentation. This gave us the complete extent of V1. For locating HVAs, we cross-checked the Fourier maps with those obtained from the responses to patches of gratings at different retinotopic locations and confirmed them using standard maps in the literature[44].

Modulation index. We calculated the modulation index as the difference between the activity during the optogenetic condition and the activity during the control condition, divided by the sum of the two.

Inclusion criteria. For significant thresholds and other details, see subsections within the 'Data analysis' section.

Figure 1b: As mentioned in 'Receptive field mapping', we estimated the centre of the receptive field by fitting the responses to patches of gratings presented along a grid with a two-dimensional Gaussian. Neurons were included if they significantly responded to patches of gratings at any location within 10° of their estimated centres, and they significantly responded to at least one inverse stimulus at any location.

Figure 1c: We estimated the centre of the receptive field by fitting the responses to patches of gratings presented along a grid with a two-dimensional Gaussian. Neurons were included if they significantly responded to patches of gratings at any location within 10° of their estimated centres, their estimated centres were within 10° of the centre of the size-tuning stimuli, and they significantly responded to at least one classical size-tuning stimulus.

Figure 1e, f: Same criteria as for Fig. 1b, c, respectively.

Figure 2: Receptive field maps and size-tuning functions, same criteria as for Fig. 1b, c, respectively.

Figure 3b–f: We estimated the centre of the receptive field by fitting the responses to patches of gratings presented along a grid with a two-dimensional Gaussian. Units were included if they significantly responded to patches of gratings at any location within 10° of their estimated centres, their estimated centres were within 10° of the centre of both the classical and inverse stimuli presented to assess the response dynamics (response dynamics stimuli), they were inverse-tuned, both their classical and inverse preferred sizes were within 10° of the size of the response dynamics stimuli, and they significantly responded to both classical and inverse response dynamics stimuli.

Figure 3g: Units were included for which the criteria used for Fig. 3b–f apply for either the classical or the inverse stimulus, but not necessarily for both. For classical stimuli: we estimated the centre of the receptive field by fitting the responses to patches of gratings presented along a grid with a two-dimensional Gaussian. Units were included if they significantly responded to patches of gratings at any location within 10° of their estimated centres, their estimated centres were within 10° of the centre of the classical stimuli presented to assess the response dynamics (classical response dynamics stimuli), their classical preferred sizes were within 10° or less than the size of the classical response dynamics stimuli, and they significantly responded to the classical response dynamics stimuli. For inverse stimuli: we estimated the centre of the receptive field by fitting the responses to patches of gratings presented along a grid with a two-dimensional Gaussian. Units were included if they significantly responded to patches of gratings at any location within 10° of their estimated centres, their estimated centres were within 10° of the centre of the inverse response dynamics stimuli, their inverse preferred sizes were within 10° or more than the size of the inverse response dynamics stimuli, and they significantly responded to the inverse response dynamics stimuli.

Figure 4: We estimated the centre of the receptive field by fitting the responses to patches of gratings presented along a grid with a two-dimensional Gaussian. Neurons were included if they significantly responded to patches of gratings at any location within 10° of their estimated centres (awake and anaesthetized), their estimated centres were within 10° of the centre of the size-tuning stimuli (awake and anaesthetized), they significantly responded to at least one classical size-tuning stimulus (awake only) and they were inverse-tuned.

Figure 5c–e: We estimated the centre of the receptive field by fitting the responses to patches of gratings presented along a grid with a two-dimensional Gaussian. Units were included if they significantly responded to patches of gratings at any location within 10° of their estimated centres, their estimated centres were within 10° of the centre of the size-tuning stimuli, they significantly responded to at least one classical size-tuning stimulus, and they were inverse-tuned.

Figure 5g: Same criteria as for Fig. 1c.

Figure 5i: We estimated the centre of the receptive field by fitting the responses to patches of gratings presented along a grid with a two-dimensional Gaussian. Boutons were included if they did not significantly respond to patches of gratings at any location within 10° of their estimated centres or if their estimated centres were not within 10° of the centre of the size-tuning stimuli, and if they significantly responded to at least one inverse size-tuning stimulus.

Statistics

We used Wilcoxon rank-sum tests for independent group comparisons, Wilcoxon signed-rank tests for paired tests and Student's *t*-tests for a single group analysis. No statistical methods were used to pre-determine sample sizes, but our sample sizes were similar to those used in previous publications. Allocation into experimental groups was not randomized. Data collection and analysis were not performed blind to the experimental conditions.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Datasets supporting the findings of this paper are available from the corresponding authors upon reasonable request.

Code availability

Custom code is available from the corresponding authors upon reasonable request.

38. Brainard, D. H. The psychophysics toolbox. *Spat. Vis.* **10**, 433–436 (1997).
39. Leinweber, M. et al. Two-photon calcium imaging in mice navigating a virtual reality environment. *J. Vis. Exp.* **84**, e50885 (2014).
40. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
41. Paxinos, G. & Franklin, K. B. J. *The Mouse Brain in Stereotaxic Coordinates* 4th edn (Academic, 2012).
42. Keller, A. J. et al. Stimulus relevance modulates contrast adaptation in visual cortex. *eLife* **6**, e21589 (2017).
43. Dombeck, D. A., Khabbaz, A. N., Collman, F., Adelman, T. L. & Tank, D. W. Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron* **56**, 43–57 (2007).

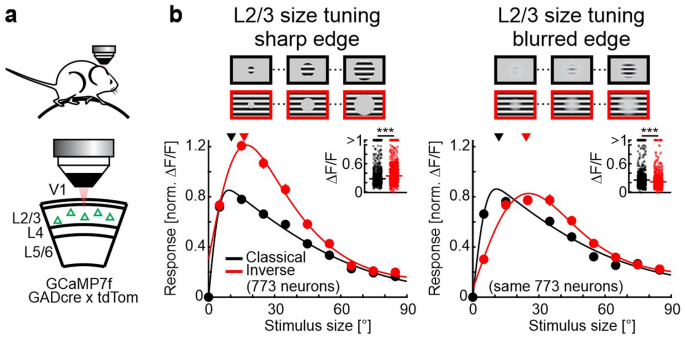
44. Garrett, M. E., Nauhaus, I., Marshel, J. H. & Callaway, E. M. Topography and areal organization of mouse visual cortex. *J. Neurosci*. **34**, 12587–12600 (2014).

Acknowledgements We thank M. Mukundan, B. Wong and L. Bao for technical support; R. Beltramo for help with extracellular recordings; J. Isaacson, G. Keller, R. Nicoll and M. Heindorf for comments on the manuscript; the members of the Scanziani laboratory for discussions of this project as well as for comments on the manuscript; and M. Rio for software support. This project was supported by the National Institutes of Health grant U19NS107613, the Howard Hughes Medical Institute and the Swiss National Science Foundation grants P300PA_177882 and P2EZP3_162284 to A.J.K. and P300PA_177898 to M.M.R. Confocal images were acquired at the Nikon Imaging Center at the University of California San Francisco.

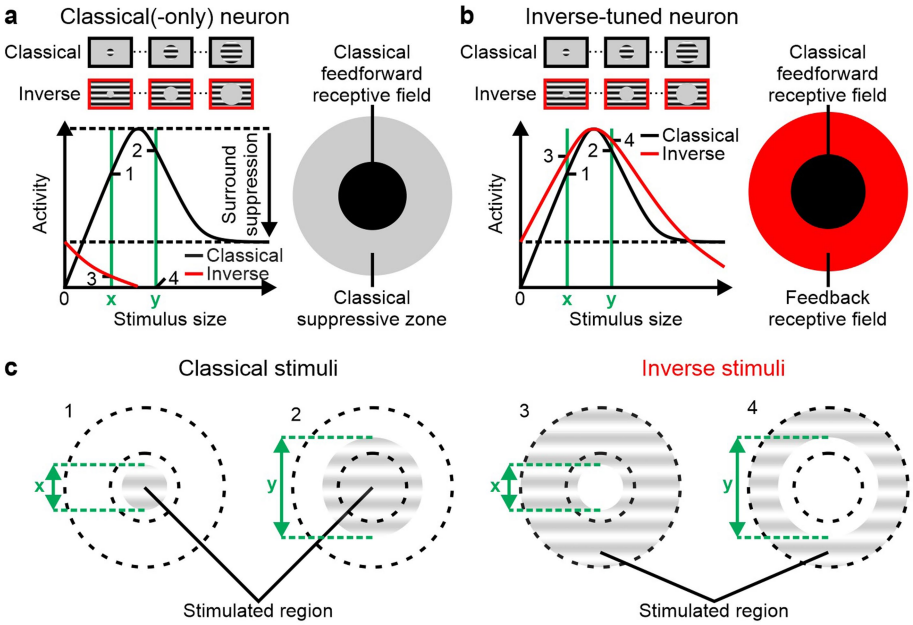
Author contributions A.J.K. and M.S. designed the study. A.J.K. and M.M.R. conducted all experiments and analysis. M.S., A.J.K. and M.M.R. wrote the manuscript.

Competing interests The authors declare no competing interests.

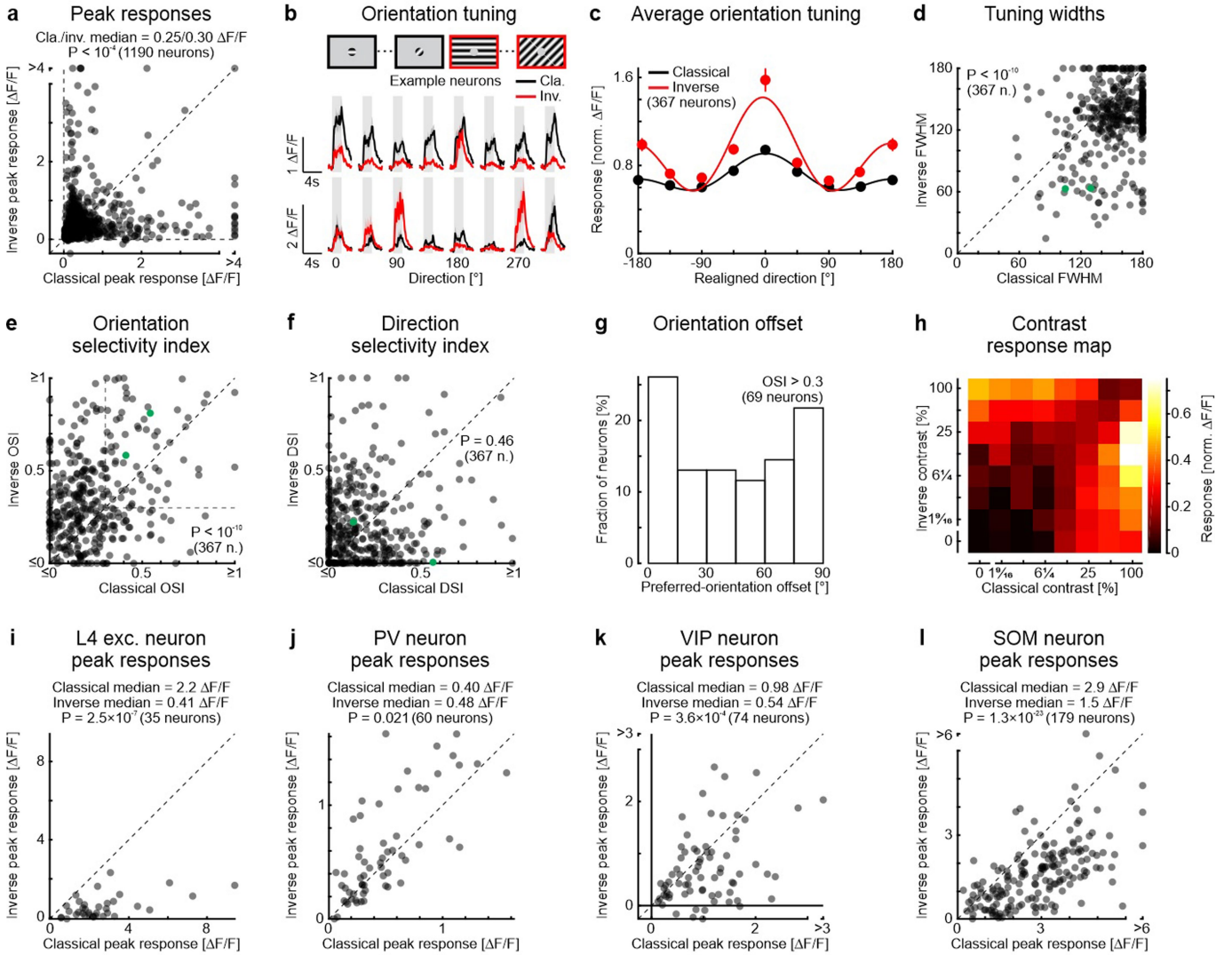
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2319-4.
Correspondence and requests for materials should be addressed to A.J.K. or M.S.
Peer review information *Nature* thanks Richard Born, Mark Hübener and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



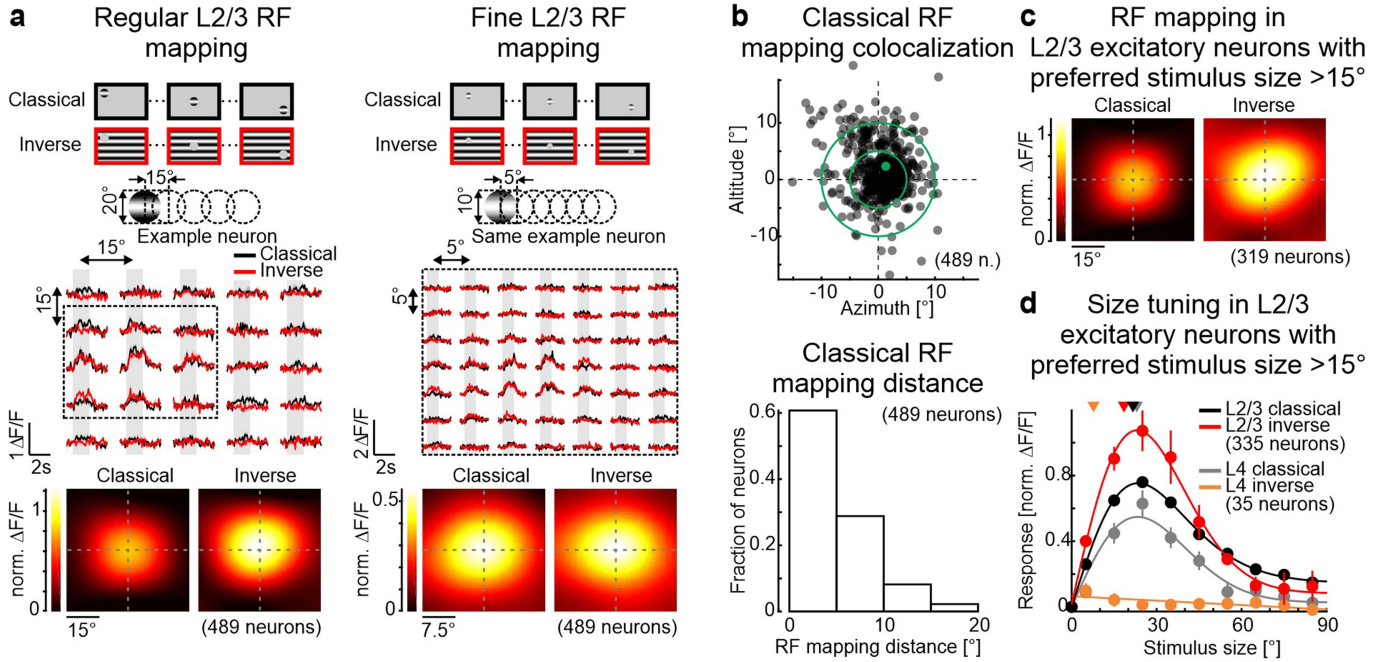
Extended Data Fig. 1 | Robust responses to inverse stimuli with blurred edges. a, Experimental configuration. **b**, Top, schematics of stimuli used for size-tuning functions. Bottom, population-averaged size tuning of classical and inverse stimuli with sharp edges (left) and blurred edges (right) (Methods). Here and in all other figures, black and red traces are responses to classical and inverse stimuli, respectively, and shaded areas are periods of stimulus presentation. Solid lines are fits to the data (Methods). Triangles above size-tuning functions indicate the median preferred size for each condition. The inset show the maximum responses, with horizontal lines indicating the medians. Two-sided Wilcoxon signed-rank test were used to compare the maximum responses for classical and inverse stimuli under each condition; sharp edge, $^{***}P = 2.0 \times 10^{-9}$; blurred edge, $^{***}P = 4.5 \times 10^{-10}$; 773 neurons in 4 mice. Data are mean (traces or data points) ± s.e.m. (shading or error bars). Here and in all other figures, error bars are present but are sometimes smaller than symbols.



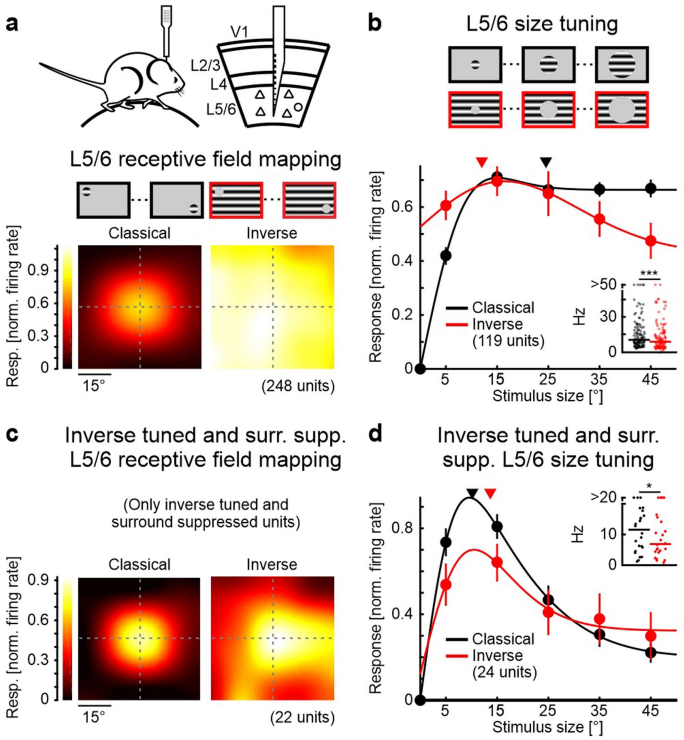
Extended Data Fig. 2 | Illustration of classical and inverse-tuned neurons. **a**, Classical(-only) neuron. Left, the response of a neuron probed with classical stimuli (black) increases with the size of the stimulus until it peaks at the preferred size of the neuron (top horizontal dotted line). The response then decreases owing to surround suppression (maximum suppressed level indicated by the lower dotted horizontal line). The response of the same neuron probed with inverse stimuli (red) starts at the maximally surround suppressed activity level (an inverse stimulus with a size of 0° is a full-field grating) and then decreases as the diameter of the grey patch increases, consistent with visual stimulation being progressively removed from the classical ffRF. Right, schematic of the ffRF of a neuron surrounded by its classical suppressive zone. **b**, Inverse-tuned neuron. Left, the response of the neuron probed with inverse stimuli (red) starts, as for the classical-only neuron, at the maximally surround suppressed activity level but then increases until it reaches the preferred inverse stimulus size of the neuron and decreases with larger diameters of the grey patch, consistent with visual stimulation being progressively removed from the fbRF. Right, schematic of the ffRF of a neuron surrounded by its fbRF. **c**, Four example stimuli: Two classical stimuli (1 and 2 of sizes *x* and *y*, respectively) and two inverse stimuli (3 and 4, also of sizes *x* and *y*, respectively). The inner dotted circle represents the outer border of the classical ffRF. The outer dotted circle represents the outer border of the suppressive region and, for inverse-tuned neurons, also the outer border of the fbRF. The response amplitudes to the four example stimuli (1 to 4) in a classical-only neuron and in an inverse-tuned neuron are marked in **a** and **b**, respectively, at the intersection between the green vertical lines (stimulus size) and the size-tuning functions.



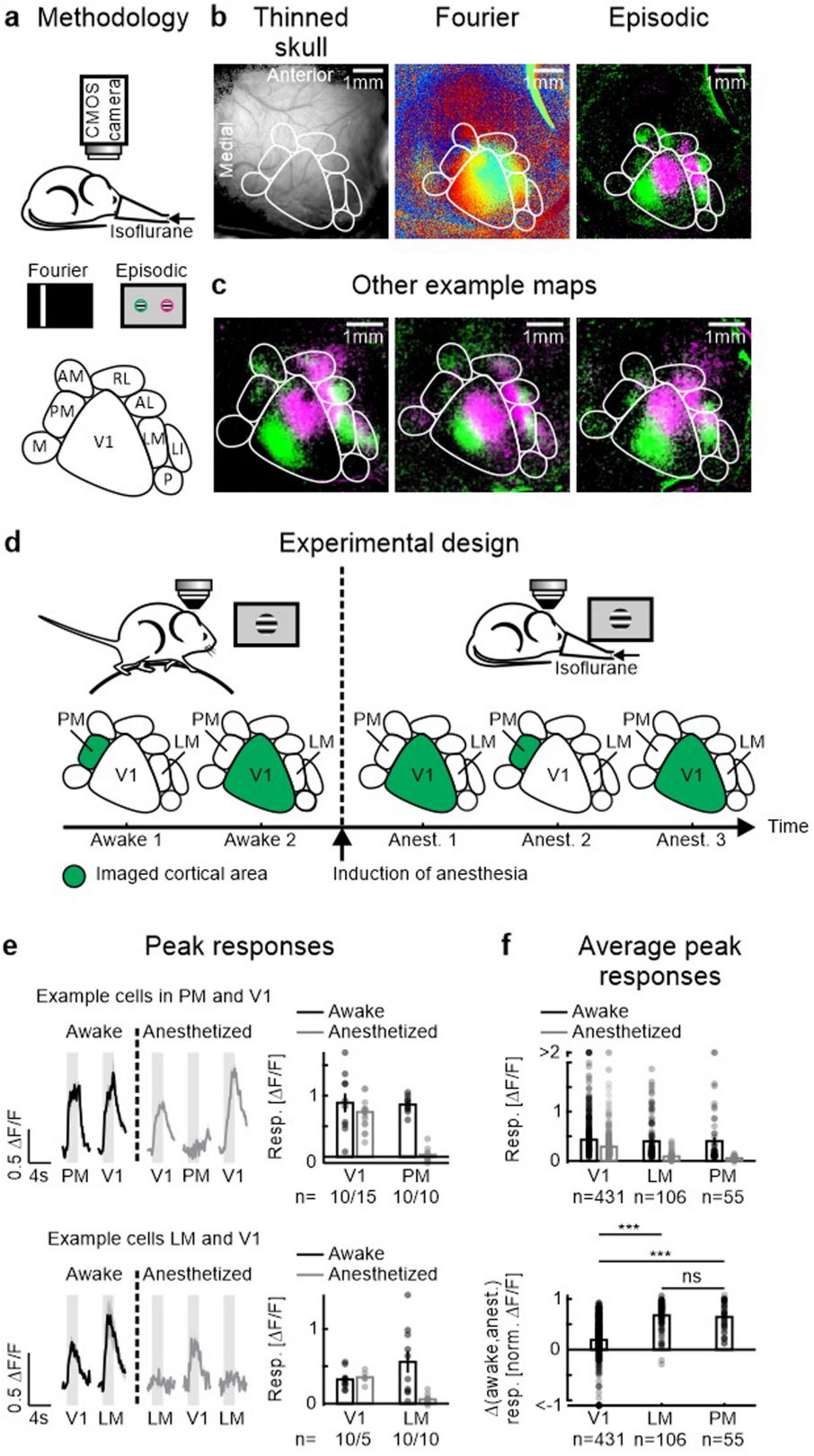
Extended Data Fig. 3 | Classical and inverse tuning properties in L2/3 excitatory neurons. a, Scatter plot of the peak responses of L2/3 excitatory neurons to classical and inverse stimuli (maximum responses to size-tuning curves in Fig. 1c). Classical and inverse median, 0.25 and 0.30 Δ*F*/*F*, respectively. Two-sided Wilcoxon signed-rank test; $P = 7.7 \times 10^{-5}$; same excitatory L2/3 neurons as in Fig. 1c; 1,190 neurons in 9 mice. **b**, Top, schematic of stimuli presented at different orientations to map the classical and inverse orientation preferences. We tested 8 orientations at intervals of 45° at the neuron's preferred stimulus size and location using either a classical or an inverse stimulus. Bottom, calcium responses of two example neurons in V1 for different orientations using classical and inverse stimuli. **c**, Population-averaged tuning curve for inverse-tuned L2/3 excitatory neurons in response to classical and inverse stimuli. The preferred orientations of each neuron (independently for classical and inverse stimuli) were aligned to 0° and its activity was normalized to its maximum response (367 neurons in 4 mice). Solid lines are fits to the data (Methods). **d**, Tuning widths of orientation tuning curves obtained with classical stimuli compared with those obtained with inverse stimuli. For each neuron, tuning width was defined as the full width at half maximum (FWHM) of the fitted tuning curve. Two-sided Wilcoxon signed-rank test; $P = 1.8 \times 10^{-21}$; same neurons as in **c**. Green symbols represent the example neurons shown in **b**. **e**, Same as **d** but for orientation selectivity indices. The horizontal and vertical lines at 0.3 delimit the orientation-selective population. Two-sided Wilcoxon signed-rank test; $P = 7.0 \times 10^{-16}$; same neurons as in **c**. **f**, Same as **d**, **e** but for direction selectivity indices. Two-sided Wilcoxon signed-rank test; $P = 0.46$; same neurons as in **c**. **g**, Distribution of orientation offsets. For orientation-selective neurons only (see **e**, with both OSIs ≥ 0.3), an orientation offset was computed, defined as the absolute difference in orientation between the preferred orientation of a neuron for a classical and an inverse stimulus. **h**, Contrast response map. Classical and inverse stimuli were presented simultaneously, and different combinations of contrasts were tested. The contrast heat map was obtained by averaging normalized activity of inverse-tuned L2/3 excitatory neurons (86 neurons in 4 mice). Data are mean (traces or data points) ± s.e.m. (shading or error bars). **i**, Scatter plot of the peak responses of L4 excitatory neurons to classical and inverse stimuli (maximum responses to size-tuning curves in Fig. 1f). Classical and inverse median, 2.2 and 0.41 Δ*F*/*F*, respectively. Two-sided Wilcoxon signed-rank test; $P = 2.5 \times 10^{-7}$; same L4 neurons as in Fig. 1f; 35 neurons in 6 mice. **j**, Scatter plot of the peak responses of PV neurons to classical and inverse stimuli (maximum responses to size-tuning curves in Fig. 2a). Classical and inverse median, 0.40 and 0.48 Δ*F*/*F*, respectively. Two-sided Wilcoxon signed-rank test; $P = 0.021$; same PV neurons as in Fig. 2a, bottom; 60 neurons in 7 mice. **k**, Scatter plot of the peak responses of VIP neurons to classical and inverse stimuli (maximum responses to size-tuning curves in Fig. 2b). Classical and inverse median, 0.98 and 0.54 Δ*F*/*F*, respectively. Two-sided Wilcoxon signed-rank test; $P = 3.6 \times 10^{-4}$; same VIP neurons as in Fig. 2b, bottom; 74 neurons in 8 mice. **l**, Scatter plot of the peak responses of SOM neurons to classical and inverse stimuli (maximum responses to size-tuning curves in Fig. 2c). Classical and inverse median, 2.9 and 1.5 Δ*F*/*F*, respectively. Two-sided Wilcoxon signed-rank test; $P = 1.3 \times 10^{-23}$; same SOM neurons as in Fig. 2c, bottom; 179 neurons in 5 mice.



Extended Data Fig. 4 | Inverse tuning is not due to low mapping resolution. **a**, Top left, schematic of regular receptive field mapping. Stimulus diameter of 20° with a grid spacing of 15°. Centre left, trial-averaged calcium responses from an example neuron for each stimulus location. Bottom left, population-averaged receptive field for responses to classical or inverse stimuli aligned to the centre of the ffRF (489 neurons in 4 mice). Right, same but for fine receptive field mapping. Stimulus diameter of 10° with a grid spacing of 5° (only for part of the visual space covered with the regular mapping, see dotted rectangle on the left). **b**, Top, spatial offset of regular ffRF mapping compared to fine ffRF mapping (same 489 neurons in 4 mice). The ffRF centre of each neuron estimated by the fine grid mapping is aligned at [0,0] and the localization of its estimated ffRF centre estimated by the regular grid is plotted with respect to the fine grid estimated centre. Bottom, distribution of distances between the centre of ffRF estimated by fine grid mapping and the centre estimated by regular grid mapping (approximately 90% of neurons have a distance between the two centres of less than 10°). The green symbol represents the example neuron shown in **a**. **c**, Population-averaged receptive field for responses to classical or inverse stimuli aligned to the centre of the ffRF and only for L2/3 neurons that had a preferred ffRF size of more than 15° (319 neurons in 9 mice). **d**, Population-averaged size-tuning functions for classical (black: L2/3 neurons with ffRF >15°, 335 neurons in 9 mice; grey: L4 neurons, 35 neurons in 6 mice) and inverse (red: L2/3 neurons with ffRF >15°, 335 neurons in 9 mice; orange: L4 neurons, 35 neurons in 6 mice) stimuli. Solid lines are fits to the data (Methods). The triangles above size-tuning functions indicate the median preferred size for each condition. Data are mean (traces or data points) ± s.e.m. (shading or error bars).

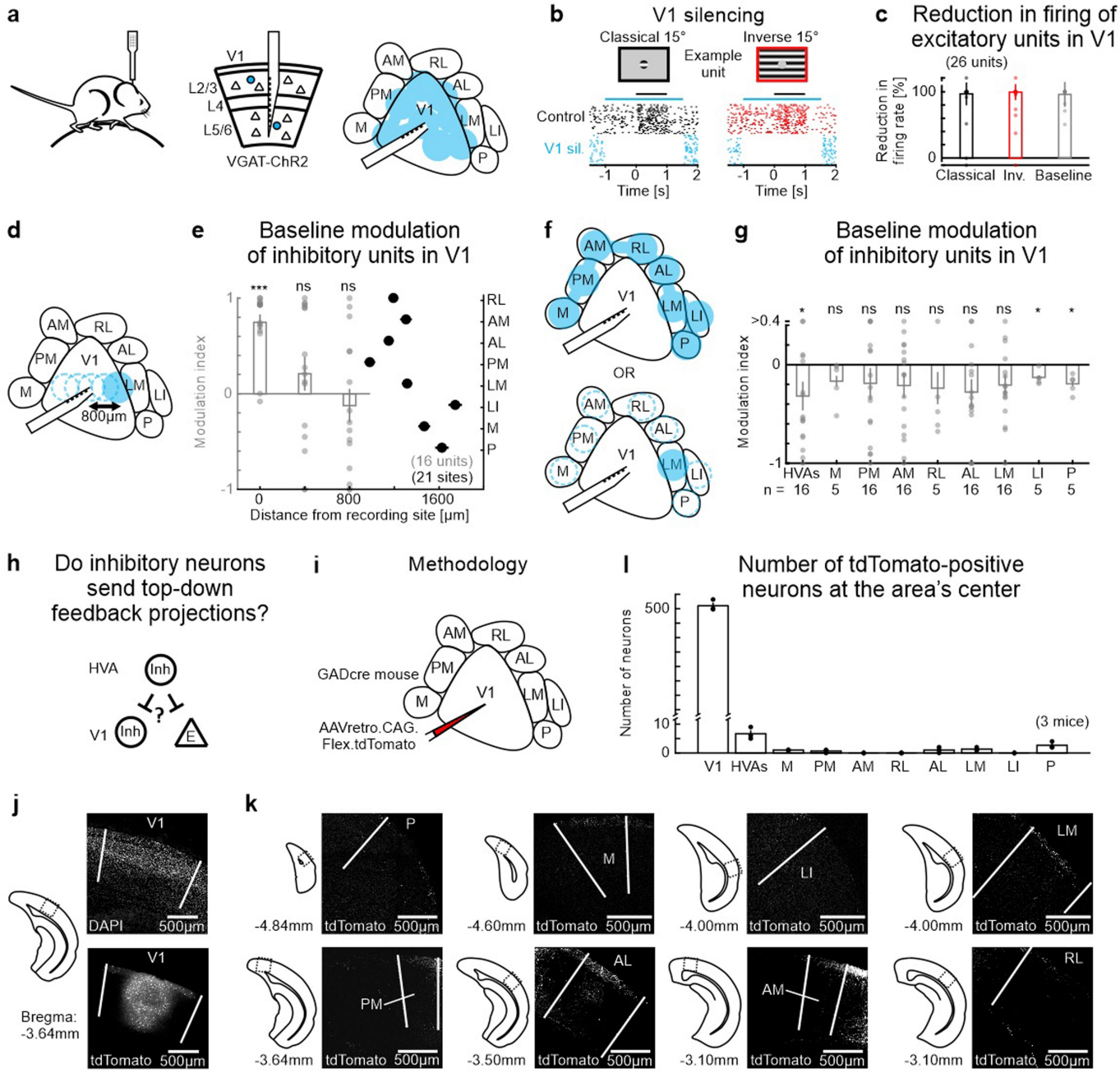


Extended Data Fig. 5 | Responses to inverse stimuli in L5/6. a, Receptive field mapping of L5/6 units using classical and inverse stimuli. Top, experimental configuration. Electrophysiological recordings were obtained in awake mice. The silicon probe spanned all layers, including deep layers (see Methods for layer definition). Centre, receptive fields were mapped using classical and inverse stimuli. Bottom left, population-averaged ffRFs for L5/6 units. Bottom right, same for inverse stimuli, aligned relative to the centre of the ffRF (248 units in 20 mice). **b**, Population-averaged size tuning of L5/6 units using classical and inverse stimuli. Top, schematic of stimuli used for size-tuning functions. The classical and inverse stimuli were presented at the same location (within 10° of the estimated centre of the ffRF). Bottom, normalized size-tuning functions for classical and inverse stimuli. Solid lines are fits to the data (Methods). Triangles above size-tuning functions indicate the median preferred size for each condition. The inset shows maximum responses, with horizontal lines denoting the median values. Two-sided Wilcoxon signed-rank test; $^{***}P = 1.1 \times 10^{-4}$; 119 units in 20 mice. **c**, **d**, Same as **a**, **b** but for a subset of L5/6 units defined both as surround-suppressed and inverse-tuned (as compared with **b**, in which all L5/6 units that responded to at least one classical stimulus size were included (Methods)); 22 units in 12 mice (**c**); Two-sided Wilcoxon signed-rank test; $^{*}P = 0.016$; 24 units in 12 mice (**d**). Data are mean ± s.e.m.



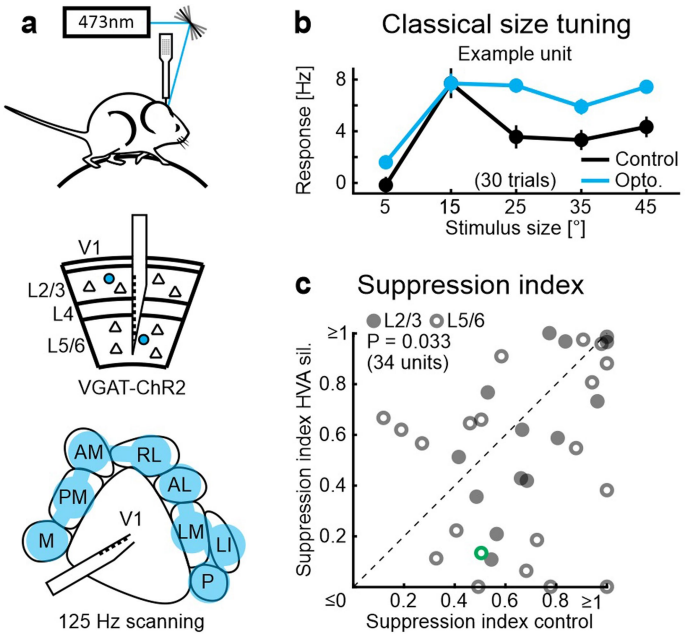
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Effect of anaesthesia is more pronounced in HVAs. **a**, Experimental configuration for intrinsic imaging of V1 and HVAs. To estimate the visual area locations and their retinotopic maps using intrinsic imaging, we presented a narrow white bar (5°) on a black background, slowly drifting (10° per second) in one of the cardinal directions ('Fourier'). We calculated the temporal phase of the Fourier component at the frequency of the bar presentation. This gave us the complete extent of V1. For locating HVAs, we cross-checked the Fourier maps with those obtained from the responses to 25° patches of gratings at different retinotopic locations ('episodic'). **b**, Left, blood vessel pattern visible through the thinned skull. Centre, Fourier map of same field of view obtained with a vertical bar moving from nasal to temporal. Right, episodic map of the same field of view. **c**, Other example episodic maps. **d**, Experimental design to assess the effect of anaesthesia on V1 and HVAs. The responses to classical stimuli of neurons in an HVA, the LM or PM, and V1 were recorded using two-photon calcium imaging. The experiment started in awake mice by imaging either an HVA or V1. After induction of anaesthesia, the same neurons were imaged again. To reduce the influence of variability in anaesthesia levels, the first imaged area under anaesthesia was imaged again at the end of the experiment. **e**, Peak responses in visual areas. Top left, example calcium response of a neuron located in PM and another neuron located in V1 in an awake mouse (black) and responses of the same neurons in the anaesthetized mouse (grey). Top right, trial-averaged peak response for the same neurons shown on the left for an awake (black) and anaesthetized (grey) mouse. Bottom, same for a different mouse but recorded in V1 and the LM. **f**, Population-averaged peak responses in awake and anaesthetized mice. Top, population-averaged peak responses in V1, the LM and PM for awake (black) and anaesthetized (grey) mice. Two-sided Wilcoxon signed-rank test; V1, $P = 6.2 \times 10^{-40}$, 431 neurons in 5 mice; LM, $P = 9.9 \times 10^{-19}$, 106 neurons in 3 mice; PM, $P = 1.1 \times 10^{-10}$; 55 neurons in 2 mice. Bottom, population-averaged difference between normalized neuronal activity for the awake and the anaesthetized state. For each neuron, all responses were normalized by the peak activity in the awake state before computing the differences. Two-sided Wilcoxon rank-sum test; V1, 431 neurons in 5 mice; LM, 106 neurons in 3 mice; PM, 55 neurons in 2 mice. Comparison of V1 and LM (V1–LM), $^{***}P = 1.2 \times 10^{-25}$; V1–PM, $^{***}P = 9.0 \times 10^{-13}$; LM–PM, NS: $P = 0.48$. Data are mean (traces or data points) ± s.e.m. (shading or error bars).

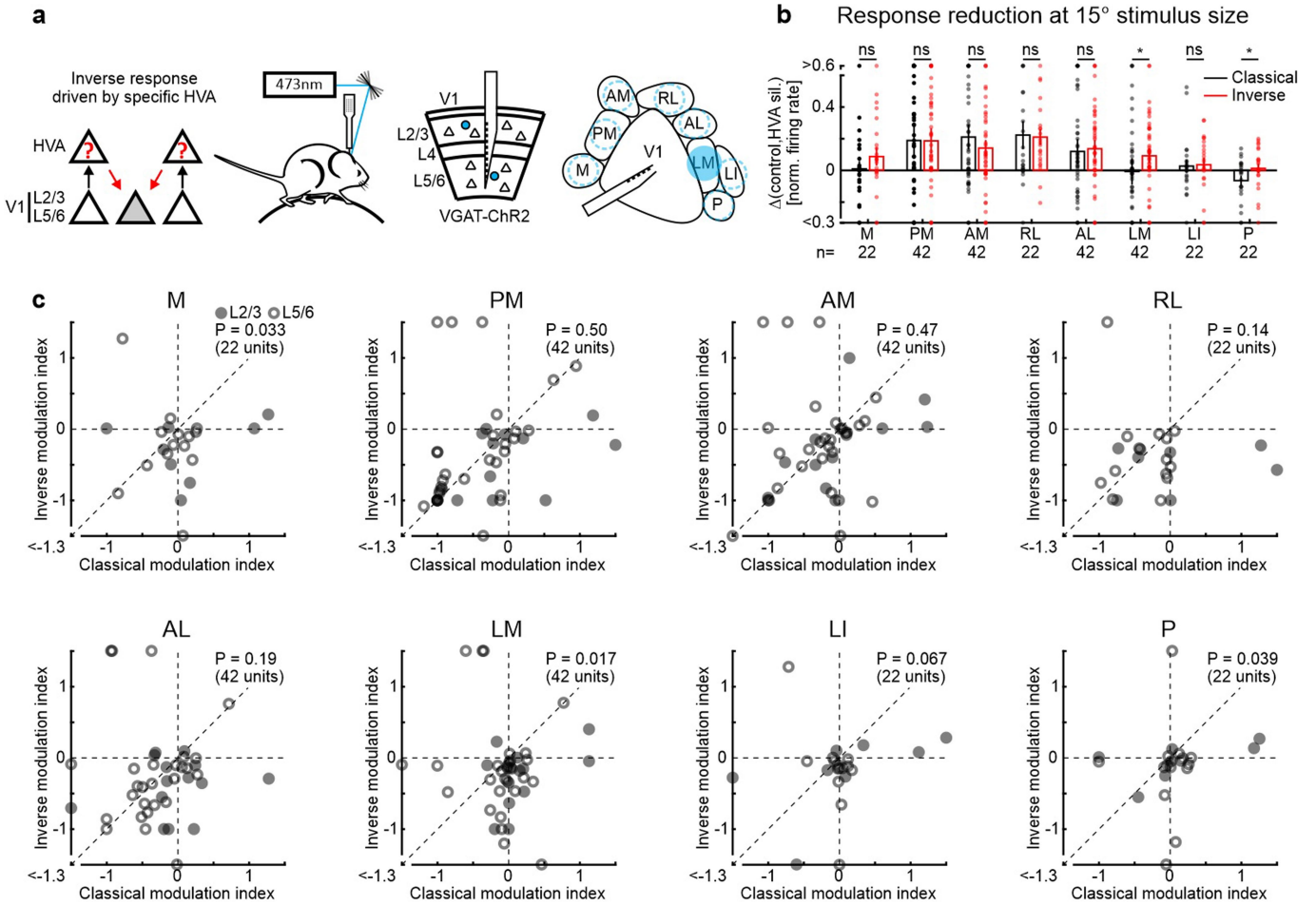


Extended Data Fig. 7 | See next page for caption.

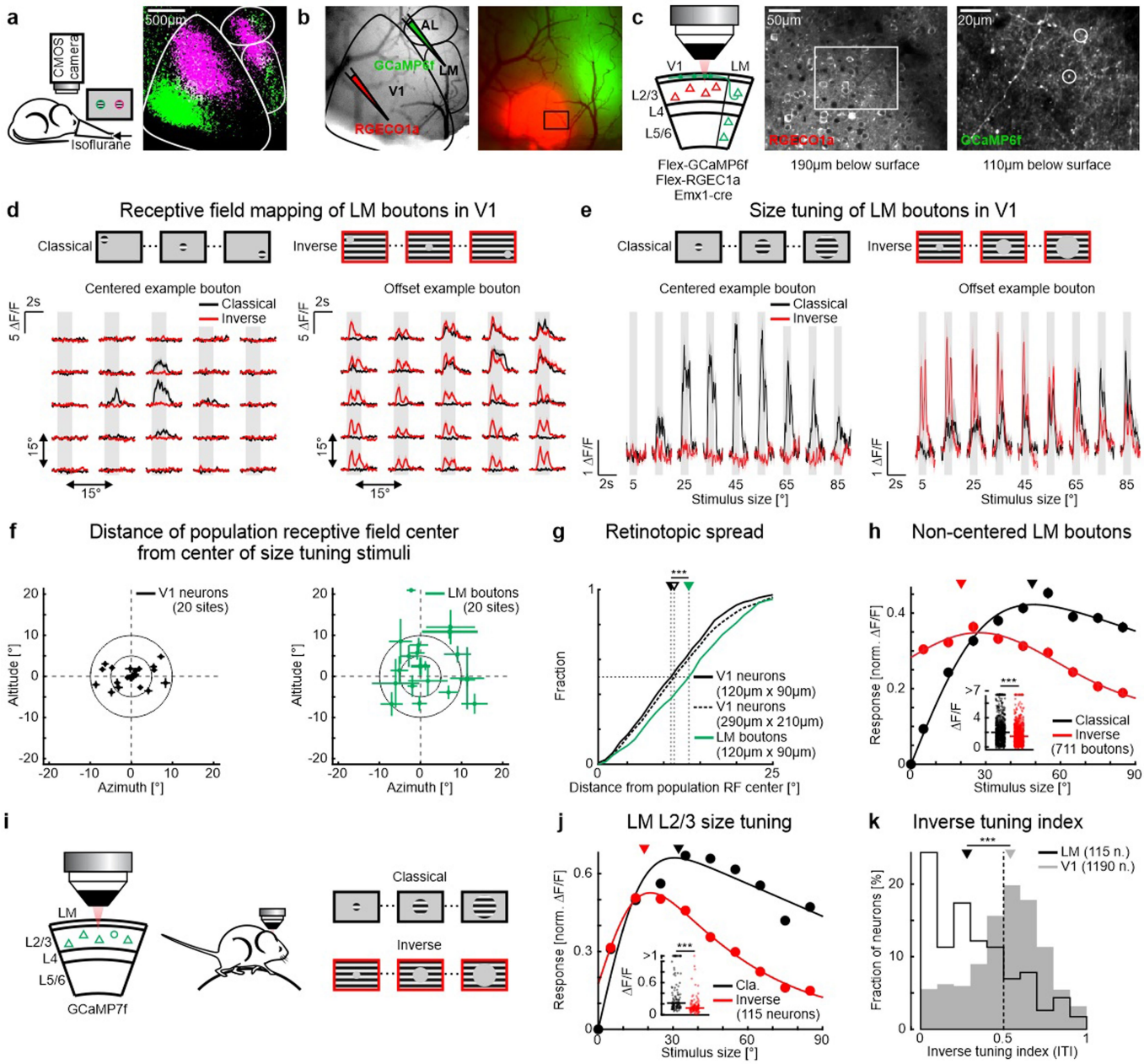
Extended Data Fig. 7 | Strong silencing by spatially restricted excitation of local inhibitory units. a, Experimental configuration. A silicon probe was inserted in V1, spanning all cortical layers, in mice expressing channelrhodopsin 2 in inhibitory neurons (VGAT-ChR2). To assess the strength of inhibition of excitatory units when using the laser scanning technique (Fig. 5, Methods), the V1 recording site as well as seven other locations were scanned at 125 Hz. **b**, Raster plot of example excitatory unit in L5/6 in response to classical and inverse stimuli of 15° diameter under control conditions (30 trials each) and during silencing of V1 (blue; V1 sil.). Black and blue horizontal lines are periods of stimulus presentation and V1 silencing, respectively. Classical and inverse stimuli were presented in random order; trials with V1 silencing were randomized as well but are separated here for clarity. **c**, Reduction in firing of excitatory units. The reduction in firing was measured as 1 – the ratio between the optogenetic condition and the control condition. Silencing reached nearly 100% for both responses to classical and inverse stimuli, and for the baseline activity (26 units in 10 mice). **d**, Experimental configuration. To assess the effect of distance on the optogenetic stimulation of inhibitory units at the recording site, two medial and two lateral locations at 400 µm and 800 µm from the V1 recording site were targeted for laser stimulation while recording in V1. **e**, Modulation of the baseline of inhibitory units. The modulation index was defined as the difference between the activity during the optogenetic and the control condition divided by the sum of the two. The modulation index was high at the recording site (at 0 µm) and quickly dropped with distance (grey bars; two-sided Student's *t*-test; 0 µm, $^{***}P = 2.0 \times 10^{-7}$; 400 µm, NS: $P = 0.26$; 800 µm, NS: $P = 0.51$; 16 units in 8 mice). As a comparison, the distance of the HVAs from the recording site is plotted on the same axis (black dots, right *y*-axis; 21 recording sites, 12 mice), suggesting that when pointing the laser at HVAs, direct activation of inhibitory neurons at the V1 recording site is unlikely. **f**, Experimental configurations. To assess the effect of the laser stimulation of HVAs on inhibitory units at the recording site, all 8 (top) or individual HVAs (bottom) were targeted for laser stimulation while recording in V1 (same configurations as during the experiments in Fig. 5 and Extended Data Figs. 8, 9). **g**, Modulation of the baseline of inhibitory units. The modulation indices were either negative or not significantly different from zero, indicating that the laser stimulation was unlikely to directly activate inhibitory neurons at the V1 recording site. Two-sided Student's *t*-test; HVA, $^{*}P = 0.045$; 16 units in 8 mice; M, NS, $P = 0.16$; 5 units in 4 mice; PM, NS, $P = 0.24$; 16 units in 8 mice; AM, NS: $P = 0.11$; 16 units in 8 mice; RL, NS: $P = 0.46$; 5 units in 4 mice; AL, NS: $P = 0.051$; 16 units in 8 mice; LM, NS: $P = 0.064$; 16 units in 8 mice; LI, $^{*}P = 0.015$; 5 units in 4 mice; P, $^{*}P = 0.010$; 5 units in 4 mice. **h**, Estimating the number of inhibitory neurons in HVA that project to V1. **i**, Methodology. A retrograde virus, AAVretro.CAG. Flex.tdTomato, was injected in V1 of GADcre mice to label glutamic acid decarboxylase (GAD)-expressing neurons projecting to the site of injection. **j**, Left, outlines of the cortical section where the confocal images shown on the right were acquired. The location of the imaged area is further indicated by the dotted square depicted on the outline. The rostro-caudal distance to bregma is indicated below the outline. Right, average intensity projection. Top right, DAPI staining highlights the higher density of neurons in L4 in V1 used to define V1 borders (white lines). Bottom right, the fluorescence of tdTomato reveals numerous cell bodies in V1 around the site of injection and even more distal in L1. **k**, Same as in **j** but only for the tdTomato fluorescence and for all HVAs targeted for laser stimulation in Fig. 5 and Extended Data Figs. 8, 9. White lines delimit the boundaries of the area. **l**, Quantification of tdTomato-positive neurons at the centre of the area. The number of tdTomato-positive neurons were counted in the section containing the centre of the investigated area. 'HVAs' represents the sum of tdTomato-positive neurons in all HVAs. Note the sparse inhibitory projections from HVAs to V1 but the abundance of local inhibitory projections within V1 (3 mice). Data are mean ± s.e.m.



Extended Data Fig. 8 | Silencing higher visual areas reduces surround suppression in V1. a, Experimental configuration. A laser beam is scanned over HVAs around V1 for optogenetic silencing while recording in V1. **b**, Size-tuning function of an example unit (baseline subtracted firing rates) to classical stimuli with (blue) or without (black) HVA silencing. Note the relief of surround suppression at larger stimulus sizes upon silencing HVAs. **c**, Scatter plot of the classical suppression index with or without silencing of HVAs (Methods). Two-sided Wilcoxon signed-rank test; $P = 0.033$; 34 units in 12 mice. Closed and open symbols are units from L2/3 and L5/6, respectively. The green symbol represents the example unit shown in **b**. Data are mean ± s.e.m.



Extended Data Fig. 9 | Silencing individual higher visual areas differentially affects responses to classical and inverse stimuli. **a**, Schematic of results and experimental configuration. Individual HVAs are targeted for optogenetic silencing while recording in V1. **b**, Difference in firing rates (baseline-subtracted and normalized) between control conditions and individual HVA silencing for classical and inverse stimuli. Two-sided Wilcoxon signed-rank test; M, NS: $P = 0.18$; 22 units in 5 mice; PM, NS: $P = 0.46$; 42 units in 12 mice; AM, NS: $P = 0.88$; 42 units in 12 mice; RL, NS: $P = 0.81$; 22 units in 5 mice; AL, NS: $P = 0.20$; 42 units in 12 mice; LM, $^*P = 0.013$; 42 units in 12 mice; LI, NS: $P = 0.51$; 22 units in 5 mice; P, $^*P = 0.020$; 22 units in 5 mice. **c**, Scatter plot of the modulation indices of individual HVA silencing for responses to classical and inverse stimuli (Methods). Closed and open symbols are units from L2/3 and L5/6, respectively. Two-sided Wilcoxon signed-rank test; M, $P = 0.033$; 22 units in 5 mice; PM, $P = 0.50$; 42 units in 12 mice; AM, $P = 0.47$; 42 units in 12 mice; RL, $P = 0.14$; 22 units in 5 mice; AL, $P = 0.19$; 42 units in 12 mice; LM, $P = 0.017$; 42 units in 12 mice; LI, $P = 0.067$; 22 units in 5 mice; P, $P = 0.039$; 22 units in 5 mice. For the visual stimulus parameters used here, the LM showed the strongest effect in preferentially reducing responses to inverse stimuli. Data are mean ± s.e.m.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Dual-colour imaging of LM boutons and their putative V1 targets. a, Left, experimental configuration. To localize V1 and the LM, we used intrinsic optical imaging (Methods). Right, response map to a nasal (magenta) and temporal patch of gratings (green). White lines represent area borders. **b**, Left, blood vessel pattern overlaid with area borders defined by the intrinsic map (black lines). The red-shifted calcium indicator RGECO1a was injected in V1 and GCaMP6f was injected in LM. Right, fluorescence of calcium indicators in V1 and the LM. The black square delimits the example imaging site shown in **c**. The scale is the same as in **a**. **c**, Left, the responses of LM boutons and of V1 cell bodies were recorded within the same cortical location. Centre, example imaging site of V1 cell bodies recorded 190 µm below the surface. The white square delimits the example imaging site shown on the right. Right, example imaging site of LM boutons in V1 recorded 110 µm below the surface. The white circles indicate the location of the example boutons in **d** and **e**. **d**, Top, schematic of receptive field mapping. Left, trial-averaged calcium responses from an example LM bouton aligned to its putative V1 target. Right, same but from an example bouton that is retinotopically offset with respect to its putative V1 target. **e**, Top, schematic of stimuli used for size-tuning functions. Left, right, trial-averaged calcium responses from the same example neurons as in **d**. **f**, Left, distance of population-averaged receptive field centre of V1 neurons from the centre of size-tuning stimuli (20 sites in 5 mice). Right, same for LM boutons. All average V1 receptive-field centres are located within 10° and average LM receptive field centres are more spread with larger standard deviations. **g**, Retinotopic spread measured as cumulative distance from population-averaged receptive-field centre. The ffRF centres of LM boutons (solid green line) were more retinotopically spread than V1 neurons measured over the same cortical surface (solid black line) or measured over approximately six times the surface of the LM bouton site (dotted black line). Two-sided Wilcoxon rank-sum test; LM–V1 same surface, $^{***}P = 1.2 \times 10^{-5}$; LM–V1 6 × surface, $^{***}P = 3.1 \times 10^{-4}$; LM, 311 boutons in 5 mice; V1 same surface, 530 neurons in 5 mice; V1 6 × surface, 2,352 neurons in 5 mice. **h**, Population-averaged size-tuning function of LM boutons (711 boutons in 5 mice) that are not retinotopically aligned with their V1 target. Both classical and inverse stimuli were presented at the ffRF location of their putative V1 targets (Methods) and not at the ffRF location of the imaged LM boutons. Solid lines are fits to the data (Methods). Triangles indicate the median preferred size. The insets display the maximum responses and horizontal lines denote the medians. Two-sided Wilcoxon signed-rank test; $^{***}P = 1.4 \times 10^{-11}$; 711 neurons in 5 mice. Data are mean ± s.e.m. **i**, Experimental configuration for two-photon calcium imaging in L2/3 neurons of the LM (green symbols) while presenting classical and inverse stimuli. **j**, Population-averaged size-tuning functions for classical and inverse stimuli. Solid lines are fits to the data (Methods). Triangles indicate the median preferred size. The insets display the maximum responses and horizontal lines denote the medians. Two-sided Wilcoxon signed-rank test; $^{***}P = 4.7 \times 10^{-10}$; 115 neurons in 3 mice. Data are mean ± s.e.m. **k**, Distribution of ITIs of LM (black) and V1 neurons (grey; same neurons as in Fig. 1c). Triangles above the distribution indicate medians. Two-sided Wilcoxon rank-sum test; $^{***}P = 2.9 \times 10^{-15}$; 115 neurons in 3 mice and 1,190 neurons in 9 mice for the LM and V1, respectively.

nature research

Corresponding author(s): Massimo Scanziani
Andreas Keller

Last updated by author(s): Feb 15, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	ScanImage (3.8 and 5.2; Vidrio); Matlab (2015a-2017a; MathWorks); Intan system and software (version 1.5; Intan Technologies); LabVIEW (2015 SP1 64 bit; National Instruments); custom code (https://sourceforge.net/projects/iris-scanning/); Psychophysics Toolbox (PTB-3; Open Source)
Data analysis	Data were analyzed with custom code written in Matlab (2015a-2017a; MathWorks) and is available online (https://sourceforge.net/projects/iris-scanning/) or from the corresponding authors on request. Electrophysiology recordings were preprocessed using KiloSort and Phy (verison 1). Some images were processed with FIJI (ImageJ 1.52n).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data and code supporting the findings of this paper are available on request from the corresponding authors.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to pre-determine sample sizes, but our sample sizes were similar to those used in previous publications.
Data exclusions	No dataset were excluded from analysis but some exclusion criteria for single neurons were applied when appropriate (e.g. when a neuron does not have its receptive field where the visual stimulus was presented, it does not make sense to include it in the further steps of the analysis). A full and detailed section is included in the Method section under 'Inclusion criteria'
Replication	Multiple experimental series using both two-photon calcium imaging and electrophysiology were performed. All attempts to replicate the data were successful.
Randomization	Allocation into experimental groups was not randomized. Mice were assigned to a specific experimental group based on their genotype and all mice types were submitted to a similar visual stimulation protocol and therefore all main findings were reproduced throughout mice genotypes.
Blinding	Data collection and analysis were not preformed blind to the experimental conditions but both collection and analysis were performed by two different experimentalists.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Mice of either sex were kept on a C57BL/6 background (except VIP-IRES-cre) and were of the following genotype: Gad2-IRES-cre (GAD2tm2(cre)Zjh ; JAX:010802) × Ai14 (Gt(ROSA)26Sortm14(CAG-tdTomato)Hze; JAX:007914) for imaging of layer 2/3 (L2/3) excitatory neurons (9 mice; Fig. 1a-c and 4, Extended Data Fig. 1, 3, and 4); Emx1-IRES-cre (Emx1tm1(cre)Krj ; JAX:005628) for imaging L2/3 excitatory neurons and axons from LM (5 mice; Fig. 5f-i, and Extended Data Fig. 10a-h); Gad2-IRES-cre (GAD2tm2(cre)Zjh ; JAX:010802) for imaging L2/3 neurons and labelling inhibitory projections (8 mice; Extended Data Fig. 6, 7h-k, and 10i-k); PV-cre (Pvalbtm1(cre)Arbr; JAX:017320) × Ai14 (Gt(ROSA)26Sortm14(CAG-tdTomato)Hze; JAX:007914) for imaging of L2/3 parvalbumin-expressing inhibitory neurons (PV; 7 mice; Fig. 2b); VIP-IRES-cre (Viptm1(cre)Zjh; JAX: 010908) × Ai14 (Gt(ROSA)26Sortm14(CAG-tdTomato)Hze; JAX:007914) for imaging of L2/3 vasoactive-intestinal-peptide-expressing inhibitory neurons (VIP; 8 mice; Fig. 2c); Sst-IRES-cre (Ssttm2.1(cre)Zjh; JAX:028864) × Ai14 (Gt(ROSA)26Sortm14(CAG-tdTomato)Hze; JAX:007914) for imaging of L2/3 somatostatin-expressing inhibitory neurons (SOM; 5 mice; Fig. 2d); Scnn1a-Tg3-cre (Tg(Scnn1a-cre)3Aibs/J; JAX:009613) and Scnn1a-Tg3-cre (Tg(Scnn1a-cre)3Aibs/J; JAX:009613) × Ai148 (Igs7tm148.1(tetO-GCaMP6f,CAG-tTA2)Hze; JAX:030328) for imaging layer 4 (L4) excitatory neurons (5 mice and 1 mouse, respectively; Fig. 1d-f, and Extended Data Fig. 4d); and VGAT-ChR2-EYFP (Tg(Slc32a1-COP4*H134R/EYFP)8Gfng/J; JAX:014548) for electrophysiology and optogenetic inhibition experiments (20 mice; Fig. 3 and 5a-e, Extended Data Fig. 5, 7a-g, 8, and 9). The mice were housed on a reverse light cycle (light/dark cycle: 12/12 hrs). At the start of the experiments, all mice were between 2 and 9 months old.
Wild animals	This study did not involve wild animals.

Field-collected samples	This study did not involve samples collected from the field.
Ethics oversight	All experimental procedures were approved by the Institutional Animal Care and Use Committee (IACUC #AN179056) of the University of California, San Francisco.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Reversing a model of Parkinson's disease with in situ converted nigral neurons

https://doi.org/10.1038/s41586-020-2388-4

Received: 12 November 2018

Accepted: 13 May 2020

Published online: 24 June 2020

Check for updates

Hao Qian[1], Xinjiang Kang[2,3], Jing Hu[1,8], Dongyang Zhang[4], Zhengyu Liang[1], Fan Meng[1], Xuan Zhang[1], Yuanchao Xue[1,9], Roy Maimon[1,5], Steven F. Dowdy[1], Neal K. Devaraj[4], Zhuan Zhou[2], William C. Mobley[6], Don W. Cleveland[1,5] & Xiang-Dong Fu[1,7]✉

Parkinson's disease is characterized by loss of dopamine neurons in the substantia nigra[1]. Similar to other major neurodegenerative disorders, there are no disease-modifying treatments for Parkinson's disease. While most treatment strategies aim to prevent neuronal loss or protect vulnerable neuronal circuits, a potential alternative is to replace lost neurons to reconstruct disrupted circuits[2]. Here we report an efficient one-step conversion of isolated mouse and human astrocytes to functional neurons by depleting the RNA-binding protein PTB (also known as PTBP1). Applying this approach to the mouse brain, we demonstrate progressive conversion of astrocytes to new neurons that innervate into and repopulate endogenous neural circuits. Astrocytes from different brain regions are converted to different neuronal subtypes. Using a chemically induced model of Parkinson's disease in mouse, we show conversion of midbrain astrocytes to dopaminergic neurons, which provide axons to reconstruct the nigrostriatal circuit. Notably, re-innervation of striatum is accompanied by restoration of dopamine levels and rescue of motor deficits. A similar reversal of disease phenotype is also accomplished by converting astrocytes to neurons using antisense oligonucleotides to transiently suppress PTB. These findings identify a potentially powerful and clinically feasible approach to treating neurodegeneration by replacing lost neurons.

Regenerative medicine holds great promise for treatment of disorders that feature loss of cells[3]. Given the plasticity of certain somatic cells[4], transdifferentiation approaches for switching cell fate in situ—thereby avoiding immune recognition—have gained momentum[2]. In the mouse brain, glial cell plasticity[5] has been leveraged to generate new neurons that lead to behavioural benefits in disease models[6,7]. However, there is limited evidence for transdifferentiated cells replacing lost neurons to reconstitute an endogenous neuronal circuit[8].

Most in vivo reprogramming relies on using lineage-specific transcription factors. We recently identified roles for the RNA-binding protein PTB and its neuronal analogue nPTB in controlling neuronal induction and maturation and demonstrated efficient conversion of mouse and human fibroblasts to functional neurons by sequential depletion of these RNA-binding proteins[9,10]. Notably, sequential downregulation of PTB and nPTB occurs naturally during neurogenesis[11], and once triggered, both PTB- and nPTB-regulated gene expression loops become self-reinforcing[9,10].

In this study, we investigate this strategy to directly convert astrocytes to dopaminergic (DA) neurons in the substantia nigra. Using a chemically induced model of Parkinson's disease in mouse, we show that dopamine neurons induced by PTB depletion potently restore striatal dopamine, reconstitute the nigrostriatal circuit, and reverse Parkinson's disease-like motor phenotypes. Given the emerging power of antisense oligonucleotides (ASOs) in modulating brain disorders[12], we also provide evidence for the use of ASOs directed against *PTBP1* (the gene that encodes PTB) as a feasible, single-step strategy for treating Parkinson's disease and perhaps other neurodegenerative diseases.

PTB- and nPTB-regulated loops in astrocytes

Astrocytes offer several advantages for in vivo reprogramming in the brain. These non-neuronal cells are abundant, proliferate upon injury, and are highly plastic with regards to cell fate[5]. As previously established in fibroblasts[9,10], PTB suppresses a neuronal induction loop in which the microRNA miR-124 inhibits the transcriptional repressor REST that suppresses many neuronal genes, including miR-124 (Fig. 1a, loop 1). Downregulation of PTB induces expression of nPTB, which suppresses the transcription activator BRN2 and the microRNA miR-9, both of which are required for neuronal maturation (Fig. 1a, loop 2). By modulating both loops, sequential downregulation of PTB and nPTB generates functional neurons from human fibroblasts[10].

[1]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA, USA. [2]State Key Laboratory of Membrane Biology and Institute of Molecular Medicine, Peking University, Beijing, China. [3]MOE Key Lab of Medical Electrophysiology, ICR, Southwest Medical University, Luzhou, China. [4]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, USA. [5]Ludwig Institute for Cancer Research, University of California, San Diego, La Jolla, CA, USA. [6]Department of Neurosciences and Center for Neural Circuits and Behavior, University of California, San Diego, La Jolla, CA, USA. [7]Institute of Genomic Medicine, University of California, San Diego, La Jolla, CA, USA. [8]Present address: Sichuan Provincial Key Laboratory for Human Disease Gene Study, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, China. [9]Present address: Key Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China. ✉e-mail: xdfu@ucsd.edu

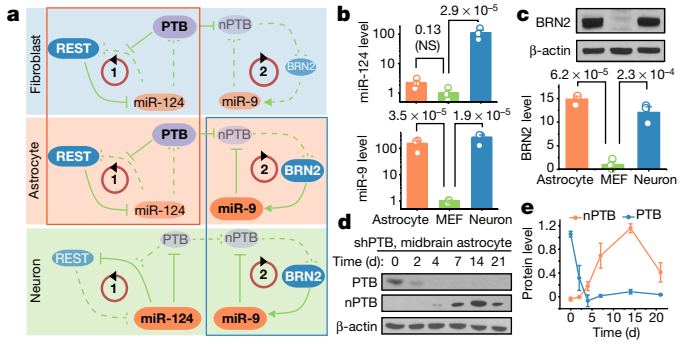


Fig. 1 | PTB knockdown induces neurogenesis in mouse and human astrocytes. a, PTB and nPTB-regulated loops critical for neuronal induction and maturation in fibroblasts, astrocytes and neurons. Bold text indicates increased expression level. The red box highlights similarity between fibroblasts and astrocytes in the PTB-regulated loop; the blue box highlights similarity between astrocytes and neurons in the nPTB-regulated loop. **b**, RT–qPCR of miR-124 and miR-9, normalized against U6 snRNA in mouse astrocytes, mouse embryonic fibroblasts (MEFs) and mouse neurons. **c**, Western blot and quantification of BRN2, normalized against β-actin in mouse astrocytes, MEFs and mouse neurons. In **b**, **c**, data are mean ± s.e.m. ($n = 3$ biological repeats); *P*-values by ANOVA with post hoc Tukey test. NS, not significant. **d**, **e**, Western blot (**d**) and quantification (**e**) of nPTB levels following PTB knockdown in mouse midbrain astrocytes. $n = 3$ biological repeats. Data are mean ± s.e.m.

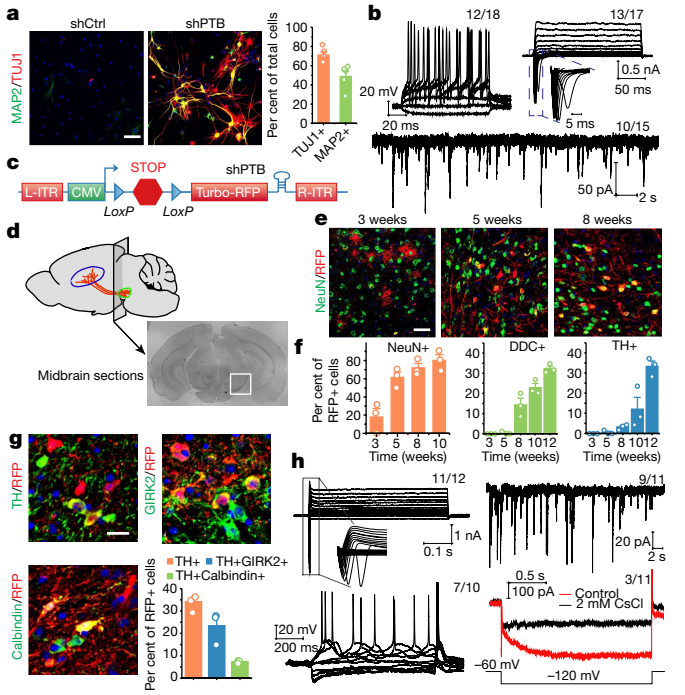


Fig. 2 | Conversion of astrocytes to functional neurons in vitro and in mouse brain. a, Left, cortical astrocytes, treated with shCtrl or shPTB lentivirus, were stained for TUJ1 (red) and MAP2 (green). Scale bar, 80 µm. Right, quantification of the numbers of cells stained with each marker ($n = 5$ biological repeats). Data are mean ± s.e.m. **b**, Electrophysiological recordings, showing repetitive action potentials (top left), large currents of voltage-dependent sodium and potassium channels (top right), and spontaneous postsynaptic currents after co-culture with rat astrocytes (bottom). Indicated in each panel is the number of cells that showed the recorded activity versus the total number of cells examined. **c**, Design of the AAV-shPTB vector. AAV-empty is the same but without shPTB. **d**, Schematic of the midbrain section used for immunochemical analysis in **e**–**h**. **e**, Gradual conversion of midbrain astrocytes to NeuN+ neurons. Representative images at three time points. Scale bar, 35 µm. **f**, Number of RFP+ cells from **e** that show positive staining for NeuN (left), DDC (middle) and TH (right). $n = 3$ biological repeats. Data are mean ± s.e.m. **g**, Converted TH+ DA neurons marked by GIRK2 or calbindin. Scale bar, 20 µm. Bottom right, results from three mice were quantified. Data are mean ± s.e.m. **h**, Electrophysiological recordings on brain slices, showing large currents from voltage-dependent sodium and potassium channels (top left), spontaneous postsynaptic currents (top right), repetitive action potentials (bottom left) and mature DA neuron-associated HCN channel activities, which are specifically blocked with 2 mM CsCl (bottom right). Indicated in each panel is the number of cells that showed the recorded activity versus the total number of cells examined.

To investigate this cascade in the conversion of astrocytes to neurons, we used mouse astrocytes from the cerebral cortex and midbrain of postnatal day (P)4 to P5 pups[13] and human fetal cortical astrocytes from gestational week 19. These cells express the astrocyte markers GFAP and ALDH1L1, but not markers for neurons and other common non-neuronal cell types in the brain (Extended Data Fig. 1a). Similar to fibroblasts, analysis by quantitative PCR with reverse transcription (RT–qPCR) showed low levels of miR-124 in the mouse and human astrocytes (Fig. 1b, Extended Data Fig. 1b). Unexpectedly, both miR-9 and BRN2 were highly expressed in astrocytes (Fig. 1b, c, Extended Data Fig. 1c). We further confirmed these expression patterns in endogenous astrocytes and neurons (Extended Data Fig. 1d). Note that expression of REST is decreased, but not eliminated, in DA neurons marked by tyrosine hydrolase (TH), consistent with its requirement for sustaining the viability of mature neurons in the brain[14]. Thus, the PTB-regulated loop in astrocytes resembles the one in fibroblasts (Fig. 1a, red box), and the nPTB-regulated loop in astrocytes resembles the one in neurons (Fig. 1a, blue box). We therefore proposed that nPTB induced by PTB knockdown would be immediately counteracted by miR-9 in astrocytes, as seen during neurogenesis from neural stem cells[15]. Indeed, unlike human dermal fibroblasts, PTB-deficient astrocytes showed transient nPTB induction (Fig. 1d, e, Extended Data Fig. 1e, f). These results suggest that astrocytes can be converted to neurons by PTB knockdown alone in both mice and humans.

Efficient astrocyte conversion in vitro

To demonstrate the functionality of converted neurons, we transduced mouse cortical astrocytes with a lentivirus expressing a small hairpin RNA (shRNA) against *Ptbp1* (shPTB). After four weeks, 50 to 80% of shPTB-transduced cells showed neuronal morphology and stained positive for the pan-neuronal markers TUJ1 and MAP2, whereas transduction with control virus did not cause expression of these markers (Fig. 2a). RNA-sequencing (RNA-seq) analysis was performed before and after conversion (Supplementary Table 1) and compared to public gene expression profiles of astrocytes and neurons (Extended Data Fig. 2a). This showed a degree of heterogeneity between independent isolates of cortical or midbrain astrocytes, but both isolates produced more homogeneous transcriptomes following conversion to neurons (Extended Data Fig. 2b, c). During conversion, typical astrocyte genes were suppressed, whereas neuronal genes were induced (Extended Data Fig. 2b, c). Notably, midbrain astrocytes gave rise to neurons expressing many DA neuron-specific genes (Extended Data Fig. 2d).

Mouse and human astrocyte-derived neurons were positive for NeuN and NSE, and most expressed markers of glutamatergic (VGlut1) or GABA (γ-aminobutyric acid)-containing (GABAergic) neurons (GAD67) (Extended Data Fig. 3a, b). Patch clamp recording six to eight weeks after conversion showed currents of voltage-gated sodium and potassium channels and repetitive action potential firing in neurons derived from both mouse and human astrocytes, and—by co-culturing the converted neurons with freshly isolated rat astrocytes—spontaneous postsynaptic events of varying frequencies were also recorded (Fig. 2b, Extended Data Fig. 3c, d). Sequential addition of antagonists of

ionotropic glutamatergic receptors (NBQX and APV) and an antagonist of $GABA_A$ receptors (picrotoxin (PiTX)) blocked the signals, indicating that the converted neurons respond to synaptic inputs from both glutamatergic and GABAergic neurons. No neuronal electrophysiological properties were detectable in astrocytes transduced with control virus (Extended Data Fig. 3e–h). These results demonstrate a one-step conversion to functional neurons by depletion of PTB.

Generating new neurons in mouse brain

We next attempted to directly reprogram astrocytes into neurons in the mouse brain. We designed an adeno-associated virus (AAV; serotype 2) vector to express shPTB (AAV-shPTB) (Fig. 2c) and a corresponding empty vector lacking shPTB (AAV-empty) as control. To enable lineage tracing, 5′ to the shPTB hairpin, we placed a red fluorescent protein (RFP)-coding sequence that was initially silenced (by a *loxP*–Stop–*loxP* cassette) but activated in cells expressing Cre recombinase. Focusing on the substantia nigra of the midbrain where DA neurons reside (Fig. 2d), we found that RFP+ cells were almost completely absent 10 weeks after injecting either AAV-empty or AAV-shPTB in wild-type mice at 1 to 2 months of age, a developmental stage when astrocytes have already lost their neurosphere-generating potential in the midbrain[16]. By contrast, RFP was expressed in response to both AAVs in *Gfap-cre* transgenic mice expressing Cre recombinase from the astrocyte-specific *Gfap* promoter[17] (Extended Data Fig. 4a, b).

Ten weeks after injection of AAV-empty into substantia nigra, most RFP+ cells were astrocytes, as indicated by typical astrocytic morphology and expression of the astrocyte markers S100b and ALDH1L1 (Extended Data Fig. 4c), with no evidence for viral transduction in NG2 cells (Extended Data Fig. 4d). We detected RFP in about 1% of NeuN+ neurons (Extended Data Fig. 4e), demonstrating minimal Cre expression in endogenous neurons in young adult mice. By contrast, 3 weeks after AAV-shPTB injection, around 20% of RFP+ cells expressed NeuN; the percentage of RFP+NeuN+ cells more than tripled by 5 weeks; and by 10 weeks around 80% of RFP+ cells were NeuN+GFAP− (Fig. 2e, f, Extended Data Fig. 4e). At this time point, most converted neurons also expressed multiple mature neuron markers (for example, MAP2, NSE and PSD95) (Extended Data Fig. 4f) and markers for glutamatergic (VGlut2) or GABAergic (GAD65) neurons (Extended Data Fig. 4g). These results demonstrate shPTB-mediated, time-dependent astrocyte-to-neuron conversion in the mouse midbrain.

Progressive maturation of new DA neurons

We next monitored the gradual appearance of DA neurons among RFP-labelled cells from 3 to 12 weeks after AAV-shPTB injection in the midbrain (Fig. 2e, f). On the basis of staining with the DA neuron markers DOPA decarboxylase (DDC) and TH, we detected a progressive increase in the number of converted DA neurons, which reached 30–35% of RFP+ cells 12 weeks after injection (Fig. 2f, Extended Data Fig. 5a, b). All RFP+TH+ DA neurons were detected proximal to, but not distal from, the site of injection where endogenous TH+RPF− DA neurons reside (Extended Data Fig. 5c–e), indicating restricted astrocyte-to-DA neuron conversion within the dopamine domain. Converted neurons also expressed multiple DA neuron markers, such as DAT, VMAT2, EN1, LMX1A and PITX3 (Extended Data Fig. 5f), with morphology similar to that of endogenous DA neurons (Extended Data Fig. 5g). A substantial population of RFP+ cells (about 22% of RFP+ cells) expressed TH and GIRK2 (a marker of A9 DA neurons and a subpopulation of A10 neurons), whereas a minor population (about 7% of RFP+ cells) expressed TH and calbindin-D28k (a marker of A10 DA neurons) (Fig. 2g), indicating that different subtypes of DA neurons were generated. Furthermore, SOX6-marked RFP+ DA neurons were confined to the substantia nigra and OTX2-marked RFP+ DA neurons were confined to the ventral tegmental area (VTA); both types expressed a common DA neuron marker, ALDH1A1 (Extended Data Fig. 5h–j). No RFP+TH+ cells were detected following injection with AAV-empty (Extended Data Fig. 5k).

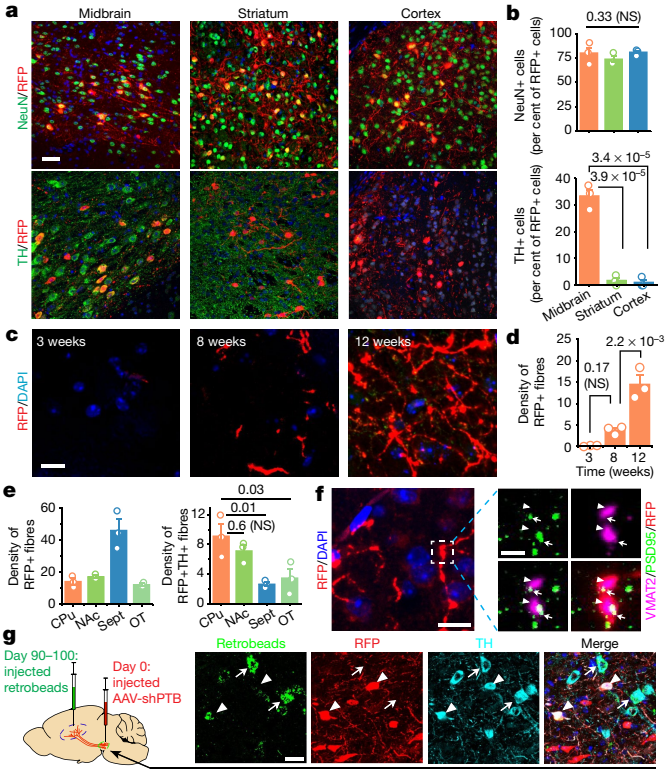


Fig. 3 | Regional specificity in astrocyte-to-neuron conversion and axonal targeting. a, **b**, Induction of NeuN+ neurons in three brain regions, showing that TH+ neurons were detected only in the midbrain. Scale bar, 40 μm. $n = 3$ mice. **c**, **d**, Progressive targeting of RFP+ fibres to striatum over the course of 12 weeks. Scale bar, 10 μm. RFP+ fibre density was determined by the sphere method and quantified (**d**) from images (**c**) at each time point. $n = 3$ biological repeats. **e**, Targeting of RFP+ fibres to multiple subregions around the striatum, particularly septal nuclei (sept) (left), but RFP+TH+ fibres are targeted mainly to CPu and NAc (right). Quantification was performed on images collected at week 12. $n = 3$ mice. **f**, Evidence for synaptic connection in CPu, as indicated in the magnified inset by colocalization of the presynaptic marker VMAT2 (arrowheads) and the postsynaptic marker PSD95 (arrows) on RFP+ fibres. Scale bar, 10 μm; magnified inset, 2 μm. **g**, Labelling of RFP+TH+ cells in substantia nigra with retrograde beads injected into striatum 90–100 days after reprogramming (left). Arrowheads indicate converted cells, one of which was labelled by the beads; arrows show endogenous (TH+RFP−) DA neurons labelled by the beads. Scale bar, 20 μm. In **b**, **d**, **e**, ANOVA with post hoc Tukey test; data are mean ± s.e.m. NS, not significant. *P*-values are indicated. In **f**, **g**, three independently repeated experiments with similar results.

Patch clamp recordings of these converted neurons (illustrated in Extended Data Fig. 6a, b) showed typical voltage-dependent currents of sodium and potassium channels, repetitive action potential firing and spontaneous postsynaptic currents. We also recorded the activity of hyperpolarization-active and cyclic nucleotide-gated (HCN) channels that could be specifically blocked with CsCl (Fig. 2h) and relatively wider action potentials compared to those of GABAergic neurons (Extended Data Fig. 6c, d)—both characteristics of mature DA neurons[18,19]. We recorded no HCN channel activities and rather infrequent firing of spontaneous action potentials at 6 weeks, and found HCN activities and increased firing of spontaneous action potentials in a fraction of RFP+TH+ DA neurons 12 weeks after injection (Extended Data Fig. 6e–g). These results demonstrate progressive functional maturation of new DA neurons within the dopaminergic neuron-containing domain of the midbrain.

Regional specificity in neuronal conversion

As controls for neuronal conversation in midbrain, we additionally injected AAV-shPTB into the cortex and striatum. While the overall conversion efficiency—based on RFP+NeuN+ cells—was similar in these three brain regions, RFP+TH+ DA neurons were detected mainly in the midbrain (Fig. 3a, b) and RFP+CTIP2+ or RFP+CUX1+ neurons were detected mostly in the cortex (Extended Data Fig. 6h). This apparent regional specificity agrees with the RNA-seq data showing that astrocytes from different brain regions exhibited different gene expression programs[20]. In our culture models, we treated cortical astrocytes with lentiviral shPTB, resulting in around 2% of cells becoming TH+ neurons, as additionally characterized by induction of the DA neuron-specific genes *Slc6a3* and *Foxa2* and positive staining for DAT, VMAT2, TH, LMX1A, PITX3 and DDC (Extended Data Fig. 7a–d). By contrast, cultured midbrain-derived astrocytes produced a fivefold higher proportion (around 10%) of TH+ neurons (Extended Data Fig. 7e–g).

We found no evidence that conditioned medium from cultured midbrain astrocytes enhanced the conversion of cortical astrocytes to TH+ neurons (Extended Data Fig. 8a, b), which prompted us to investigate other potential cell-autonomous contributions to the regional specificity by performing RT–qPCR analysis on isolated cortical and midbrain astrocytes. Relative to cortical astrocytes, midbrain astrocytes expressed higher basal levels of transcription factors enriched in DA neurons (Extended Data Fig. 8c, d) and, in response to PTB depletion, these transcription factors were more robustly induced in midbrain astrocytes relative to cortical astrocytes (Extended Data Fig. 8e, f). These findings suggest that distinct promoter–enhancer networks may underlie the regional specificity for astrocytes from different brain regions, as recently observed in microglia[21]. The higher DA neuron conversion rate also enabled us to record dopamine release from midbrain astrocyte-derived neurons (Extended Data Fig. 8g–i). These in vitro studies strongly suggest that higher basal levels and more robust induction of lineage-specific transcription factors may contribute to the higher propensity of midbrain astrocytes to generate DA neurons. The much higher conversion efficiency in the mouse midbrain (about 35%) compared with isolated midbrain astrocytes (about 10%) also points to the contribution of the local microenvironment to DA neuron conversion from midbrain astrocytes.

Innervation in the nigrostriatal pathway

We next investigated the dynamics of fibre outgrowth from newly converted neurons in the brain. We initially monitored the outgrowth of RFP+ fibres along the nigrostriatal bundle (Extended Data Fig. 9a, b). Using the sphere method[22], we quantified the fibre density, revealing a time-dependent appearance of RFP+ fibres in the nigrostriatal bundle, reaching 29.6 ± 5.4 fibres by 12 weeks, with 5.75 ± 0.5 fibres being RFP+TH+ (Extended Data Fig. 9c, d) (mean ± s.e.m.). As DA neurons usually target striatum, we also detected progressively increasing numbers of RFP+ fibres in this distal region, reaching 14.5 ± 3.6 fibres per area by 12 weeks (Fig. 3c, d). Examining brain regions more broadly, we found that RFP+ fibres targeted caudate putamen (CPu) as well as nucleus accumbens (NAc), septal nuclei and olfactory tubercle (Extended Data Fig. 9e), as previously observed with grafted neuronal stem cells[22]. A fraction of these RFP+ fibres were also TH+ (Extended Data Fig. 9f). Of note, despite around threefold more RFP+ fibres in septal nuclei, RFP+TH+ processes were about fourfold more abundant in both CPu and NAc regions (Fig. 3e). Focusing on the CPu, we detected colocalization of the presynaptic marker VMAT2 and the postsynaptic marker PSD95 on RFP+ fibres, suggesting the presence of synaptic connections (Fig. 3f).

To further substantiate functional targeting to striatum, we injected green fluorescent retrobeads into the CPu region of mice 1 month or 3 months after AAV-shPTB delivery to enable axonal uptake and retrograde labelling of the corresponding cell bodies (Fig. 3g, left). One day after injection, we saw green retrobeads in both endogenous TH+RFP− cells and converted TH+RFP+ cells in the substantia nigra. We could detect labelling of only endogenous DA neurons after 1 month following AAV-shPTB transduction (Extended Data Fig. 9g, h), and after 3 months, we detected retrobeads in both endogenous (RFP−) and newly converted (RFP+TH+) neurons (Fig. 3g). These results demonstrate time-dependent incorporation of new DA neurons into the nigrostriatal pathway.

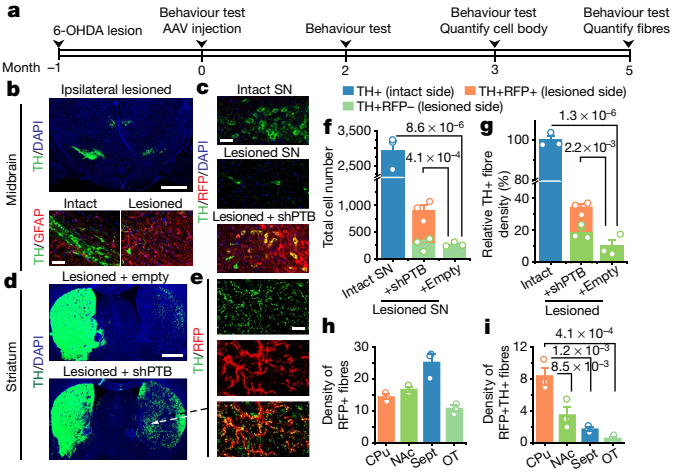


Fig. 4 | Replenishing lost DA neurons to reverse parkinsonian phenotype. **a**, Schematic of the experimental schedule for 6-OHDA-induced lesion in substantia nigra (SN) followed by AAV-shPTB treatment and behavioural tests. **b**, Unilateral loss of TH+ cells in midbrain induced by 6-OHDA (top; scale bar, 500 µm) with increased numbers of GFAP+ astrocytes (bottom; scale bar, 50 µm). **c**, Comparison between unlesioned (top) and 6-OHDA-lesioned substantia nigra (middle), showing converted DA neurons (yellow) after AAV-shPTB treatment (bottom). Scale bar, 50 µm. **d**, **e**, TH+ fibres in striatum treated with AAV-empty (top) or AAV-shPTB (bottom). Scale bar, 500 µm. **e**, Magnified views from **d**, showing extensive RFP+TH+ fibres. Scale bar, 10 µm. In **b**–**e**, three independently repeated experiments with similar results. **f**, **g**, Quantification of cell bodies (**f**) and fibres (**g**) in DA neurons in the unlesioned side (blue), remaining endogenous RFP−TH+ DA neurons in the lesioned side (green), and converted RFP+TH+ DA neurons in the lesioned side (orange). Data were from two sets of mice ($n = 3$ in each set) transduced with AAV-shPTB or AAV-empty. **h**, **i**, Quantification of RFP+ (**h**) or RFP+TH+ (**i**) fibre density in the indicated subregions of the brain ($n = 3$ mice in each group). OT, olfactory tubercle. In **f**, **g**, **i**, ANOVA with post hoc Tukey test; data are mean ± s.e.m. *P*-values are indicated.

Replenishing lost DA neurons in a disease model

Following the successful generation of DA neurons, we investigated their potential to reconstitute an injured nigrostriatal pathway. We selected a widely used model of Parkinson's disease in mouse, in which DA neurons are efficiently ablated by 6-hydroxydopamine (6-OHDA), a dopamine analogue that is toxic to DA neurons[23]. Although this model does not recapitulate all essential features of Parkinson's disease pathogenesis[24], it does result in a critical endpoint—the loss of neurons in the substantia nigra and depletion of striatal dopamine. One month after 6-OHDA injection into one side of the medial forebrain bundle (Fig. 4a), we observed unilateral loss of TH+ cell bodies in the midbrain (Fig. 4b, top), accompanied by a marked increase in GFAP+ astrocytes (Fig. 4b, bottom), indicative of the expected astrocytic response[25]. One month after the lesion, we injected AAV-empty or AAV-shPTB in the lesioned side and observed increased RFP+TH+ cell bodies around 10–12 weeks later with AAV-shPTB, but not with AAV-empty (Fig. 4c, Extended Data Fig. 10). We also detected a marked increase in RFP+TH+ fibres in striatum of mice treated with AAV-shPTB, but not in those treated with AAV-empty (Fig. 4d, e, Extended Data Fig. 11a, b).

Quantitative analysis revealed that the initial 2,926 ± 273 TH+ neuronal cell bodies in substantia nigra were reduced by around 90% (to 266 ± 22) following the lesion and AAV-shPTB induced 634 ± 38 new RFP+TH+ neurons (Fig. 4f), thereby restoring TH+ neurons to approximately one third (904 ± 108) of the initial number. Similarly, 6-OHDA lesioning reduced the number of TH+ fibres by around 90% and AAV-shPTB restored total TH+ fibre density to about 30% of the density in the uninjured brain (Fig. 4g). We detected a slight increase in TH+RFP− fibre density following treatment with AAV-shPTB compared with AAV-empty (Extended Data Fig. 11c, d), suggesting that AAV-shPTB treatment might aid recovery of some remaining damaged endogenous DA neurons. Quantification of total RFP+ fibres versus RFP+TH+ fibres in different striatal regions and surrounding areas revealed that while the septal nuclei was enriched with RFP+ fibres (Fig. 4h), the CPu contained the highest proportion of RFP+TH+ fibres (Fig. 4i, Extended Data Fig. 12). Thus, without additional treatment to specify neuronal subtypes, AAV-shPTB is sufficient to induce new DA neurons from endogenous midbrain astrocytes that partially restore lost DA neurons and their axons within the nigrostriatal dopamine pathway.

Restoration of striatal dopamine

We next investigated whether AAV-shPTB-induced neurons would restore dopamine levels in the striatum by preparing extracts and quantifying dopamine levels using high-performance liquid chromatography (HPLC) (Fig. 5a). Samples were spiked with known quantities of dopamine to define the elution position and to establish the relationship between signal intensities and the amount of dopamine (Extended Data Fig. 13a, b). We detected similar amounts of dopamine in both sides of uninjured mice (Fig. 5b) and found that 6-OHDA lesion reduced dopamine to about 25% of the normal level (Fig. 5c). Treatment with AAV-shPTB, but not with AAV-empty, markedly increased the dopamine level compared with lesioned striatum (Fig. 5d), reaching approximately 65% of the uninjured level (Fig. 5e).

To test whether DA neuron function was restored, we directly measured activity-induced dopamine release to demonstrate restored DA neuron functions by inserting a stimulating electrode in the medial forebrain bundle and a carbon fibre electrode in striatum of live mice (Fig. 5f). In lesioned mice treated with AAV-empty, we recorded stimulation-dependent dopamine release in the uninjured side but a greatly diminished signal in the lesioned side (Fig. 5g, left). In lesioned mice treated with AAV-shPTB, activity-induced dopamine release was detected in both the uninjured and lesioned sides (Fig. 5g, right). Three out of four mice showed significant restoration of dopamine release (Fig. 5h). Placing a stimulating electrode and carbon fibre electrode on striatal slices from the same group of mice (Fig. 5i), we recorded activity-induced dopamine release (Fig. 5j), with the same mouse showing reduced release as in live recording (Fig. 5k), ruling out a misplaced electrode as a cause of reduced release in vivo. These data demonstrate robust restoration of striatal dopamine and activity-induced dopamine release in AAV-shPTB-reprogrammed mice.

Reversing disease-relevant motor phenotypes

Next, we tested the ability of AAV-shPTB to restore motor function to mice with 6-OHDA lesions. We performed three common behavioural tests, two based on drug-induced rotation and one based on spontaneous motor activities[26]. Contralateral rotation induced by apomorphine and ipsilateral rotation triggered by amphetamine were markedly increased following lesion with 6-OHDA; both phenotypes were progressively restored to nearly wild-type levels within three months after AAV-shPTB treatment (Fig. 6a, b). No correction was recorded in mice treated with AAV-empty (Fig. 6a) or with non-specific AAV-shGFP (Extended Data Fig. 13c).

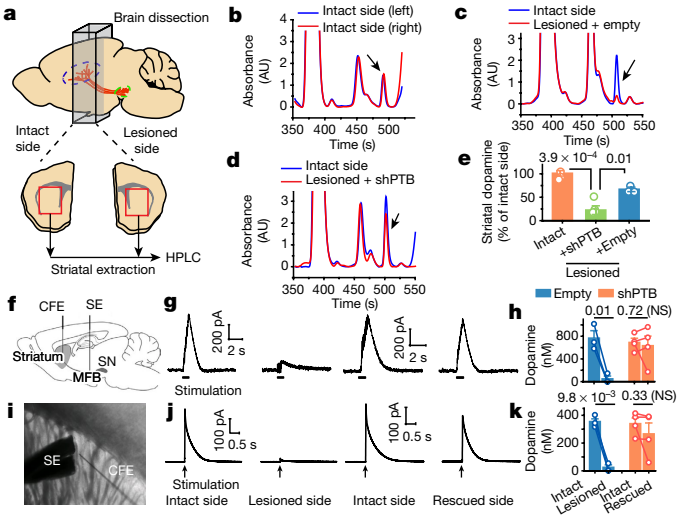


Fig. 5 | Restoration of dopamine biogenesis and activity-induced dopamine release. a, Schematic depiction of the measurement of striatal dopamine levels by HPLC. **b**–**d**, Striatal dopamine levels in two sides of unlesioned mouse brain (**b**), comparison between unlesioned and 6-OHDA-lesioned sides (**c**) and restoration in the lesioned side after reprogramming in ipsilateral nigra (**d**). Arrows in each panel indicate the position of dopamine in the HPLC profile. AU, absorbance units. **e**, Striatal dopamine restoration after reprogramming with AAV-shPTB in ipsilateral substantia nigra ($n = 3$ unlesioned mice or lesioned mice treated with AAV-shPTB; $n = 4$ lesioned mice treated with AAV-empty). **f**–**h**, Activity-induced dopamine release in striatum. **f**, Schematic of striatal dopamine recording with insertion of a carbon fibre electrode (CFE) in striatum and stimulation electrode (SE) in the medial forebrain bundle (MFB) next to the substantia nigra in live mice. **g**, Representative traces of activity-induced dopamine release recorded on the unlesioned and 6-OHDA-lesioned striatum before and after neuronal conversion. **h**, Overall recorded results ($n = 3$ for AAV-empty; $n = 4$ for AAV-shPTB). Circles represent individual mice; lines connect recordings from the same mice before and after reprogramming. **i**–**k**, Dopamine release in striatal slices from the same set of mice analysed in **g**, as shown in **i**. Representative traces (**j**) and overall results (**k**). In **e**, ANOVA with post hoc Tukey test; in **h**, **k**, Student's *t*-test; data are mean ± s.e.m. *P*-values are indicated.

To examine spontaneous motor activity, we scored limb-use bias. Uninjured mice used both limbs with relatively equal frequency, whereas unilaterally lesioned mice showed preferential ipsilateral touches, indicating disabled contralateral forelimb function. In lesioned mice transduced with AAV-shPTB, we observed a time-dependent improvement in contralateral forelimb use, whereas mice transduced with AAV-empty did not show any improvement (Fig. 6c). These results demonstrate essentially full correction of the motor phenotypes in this chemically induced model of Parkinson's disease. As Parkinson's disease and most other types of neurodegenerative diseases show age-dependent onset, we extended our approach from relatively young (two-month-old) mice to one-year-old mice, an age comparable to the age of onset of Parkinson's disease in humans. Of note, while the behavioural benefits of AAV-shPTB transduction following apomorphine-induced rotation did not reach statistical significance—perhaps owing to relatively unstable phenotype scored by this assay in aged animals (Fig. 6d)—a substantial improvement was recorded with the limb-use asymmetry test (Fig. 6e). These observations point to an age-related decrease in neuronal reprogramming, a critical challenge to be met in future studies.

Chemogenetic analysis of new DA neurons

We used the DREADD platform[27]—a chemogenetic approach—to test whether new DA neurons are directly responsible for the restoration

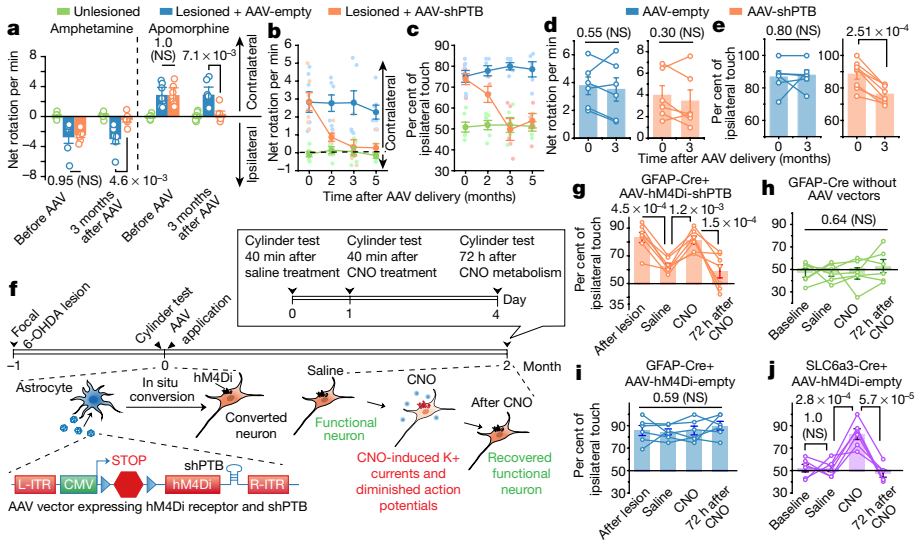


Fig. 6 | Behavioural benefits and chemical genetic evidence for induced neurons in brain repair. a, Behavioural study of mock-treated (green) or 6-OHDA-lesioned mice treated with AAV-empty (blue) or AAV-shPTB (orange). Rotation was induced by amphetamine (left) or apomorphine (right). $n = 7$ mice used for lesioned and treated conditions with AAV-empty or AAV-shPTB in apomorphine test; $n = 6$ mice for the other conditions. **b**, **c**, Time-course analysis of behavioural recovery. Rotation induced by apomorphine (**b**) and cylinder test for ipsilateral touches (**c**) in unilaterally lesioned mice ($n = 6$ or 7 mice analysed in each group, as in **a**). Data are mean ± s.e.m. **d**, **e**, Apomorphine-induced rotation test (**d**) and cylinder test (**e**) on 1-year-old lesioned mice 3 months after treatment with AAV-empty or AAV-shPTB. $n = 8$ mice used for lesioned and treated conditions with AAV-shPTB in cylinder test; $n = 6$ mice for the other conditions. Circles represent individual mice; lines connect data from the same mouse before and after reprogramming. **f**, Schematic of the chemogenetic strategy to demonstrate that converted neurons are directly responsible for phenotypic recovery, emphasizing the rapid effect of injected CNO in inhibiting activities of reprogrammed neurons and reappearance of Parkinson's disease-like phenotype after CNO metabolism. **g**–**i**, Results of cylinder test before and after injecting AAV-hM4Di-shPTB and treatment with saline or CNO or 3 days after drug withdrawal (**g**) in unlesioned mice mock-treated or treated with CNO (**h**) or lesioned mice treated with AAV-hM4Di-empty vector (**i**). $n = 7$ *Gfap-cre* mice for **g**; $n = 6$ for **h**, **i**. **j**, The cylinder test results on unlesioned SLC6a3 transgenic mice ($n = 6$) treated with AAV-hM4Di, which specifically targets endogenous DA neurons owing to DA neuron-restricted Cre expression via the *Slc6a3* promoter. Statistical significance was determined by ANOVA with post hoc Tukey test (**a**, **g**–**j**) or Student's *t*-test (**d**, **e**) and is presented as mean ± s.e.m. *P*-values are indicated.

of motor function (Fig. 6f). We replaced RFP in our AAV-shPTB vector with a gene encoding an engineered inhibitory muscarinic receptor variant hM4Di, which responds to clozapine-*N*-oxide (CNO) instead of acetylcholine[28]. As with the original AAV-shPTB, expression of both hM4Di and shPTB was activated in astrocytes in *Gfap-cre* mice. Neurons converted from astrocytes would be expected to incorporate this receptor into their plasma membrane and respond to CNO to activate G_i signalling, leading to hyperpolarization and suppression of electrical activity[29]. CNO is metabolized 2 to 3 days after administration to allow functional restoration of hM4Di-expressing neurons[29].

As expected, the limb-use bias test indicated that motor performance of *Gfap-cre* mice with 6-OHDA lesions was restored two months after AAV-hM4Di-shPTB transduction. The lesion-induced phenotype reappeared within 40 min of intraperitoneal injection of CNO, but not with saline; moreover, the CNO-provoked motor phenotype disappeared within three days (Fig. 6g). CNO injection into uninjured mice had no effect, indicating that the drug did not affect endogenous DA neurons (Fig. 6h). AAV-hM4Di-empty showed no benefit to mice with lesions and had no effect on uninjured mice with or without CNO treatment (Fig. 6i, Extended Data Fig. 13d), demonstrating that the improvement in behaviour observed with AAV-hM4Di-shPTB was dependent on reprogramming.

Targeted expression of hM4Di in endogenous DA neurons (by injecting AAV-hM4Di-empty into the midbrain of mice expressing Cre from the DA neuron-specific *Slc6a3* gene promoter) was sufficient to induce the Parkinson's disease-like phenotype, but only in the presence of CNO (Fig. 6j), indicating that the introduction of the receptor into endogenous DA neurons had the intended CNO-mediated inactivating effect. Collectively, these data provide strong evidence that activity-induced signalling by astrocyte-derived neurons is responsible for phenotypic recovery.

ASO-based neuronal conversion and rescue

The PTB-regulatory loop is self-reinforcing once it is triggered by initial PTB knockdown (Fig. 1a). In response to a reduction in PTB, miR-124 becomes more efficient at targeting *REST* (owing to the ability of PTB to directly compete with the miRNA-targeting site in the 3′-untranslated region of the mRNA encoded by *REST*), resulting in decreased levels of REST, which drives further de-repression of miR-124 and thus further suppression of PTB[9,10]. This suggested that transient suppression of PTB might be sufficient to generate new neurons through ASO-mediated PTB mRNA degradation by intranuclear RNase H[12]. We thus synthesized and screened PTB ASOs containing a phosphorothioate backbone[12] and a 3′ fluorescein. An ASO targeting GFP served as control. Three PTB ASOs, but not GFP ASO, reduced PTB expression upon transfection into mouse astrocytes (Fig. 7a). The best targeting PTB ASO (no. 4), but not GFP ASO, induced expression of multiple neuronal markers, including TUJ1, MAP2, NSE and NeuN, after 5 weeks (Fig. 7b). A fraction of converted neurons were dopaminergic, as indicated by TH staining (Fig. 7b). Patch clamp recording showed that these in vitro-converted neurons were functional (Extended Data Fig. 14a–c).

We next injected PTB ASO or control GFP ASO into the midbrain of transgenic mice carrying a tamoxifen-inducible Cre that was selectively expressed in astrocytes and a tdTomato-encoding gene activated by Cre (Extended Data Fig. 14d, e). We induced Cre in these mice at P35 and stereotactically injected ASOs unilaterally into their substantia nigra 2 weeks later. PTB ASO converted a fraction of tdTomato-labelled cells to NeuN+ neurons after 8 weeks (Fig. 7c) and to TH+ neurons after

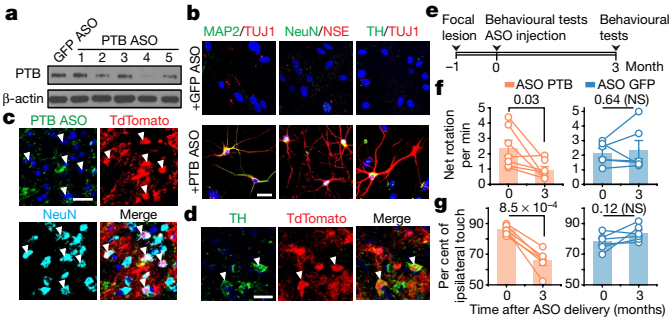


Fig. 7 | Proof-of-concept experiments with the ASO-based strategy. **a**, Screening for PTB ASOs by western blotting in mouse astrocytes. **b**, Neurons in isolated mouse cortical astrocytes induced with PTB ASO in vitro (**b**) stained for TUJ1 and MAP2 (left), NSE and NeuN (middle); a small fraction of converted neurons stained positively for TH (right). In **a**, **b**, $n = 3$ biological repeats. Scale bar, 20 µm. **c**, **d**, A proportion of tdTomato-labelled cells became NeuN+ by 8 weeks (**c**) and TH+ by 12 weeks (**d**) after injection of PTB ASO into the midbrain of *Gfap-creER*^{TM};*Rosa-tdTomato* transgenic mice. In **c**, **d**, $n = 4$ biological repeats. Scale bar, 20 µm. **e**–**g**, Schematic of 6-OHDA induced lesion, ASO treatment and behavioural tests (**e**) and results of apomorphine-induced rotation (**f**) and cylinder (**g**) tests. Circles represent individual mice; lines connect readings from the same mice before and after reprogramming ($n = 7$ used for lesioned and treatment with PTB ASO in apomorphine test; $n = 6$ for the other conditions; wild-type C57BL/6 mice). In **f**, **g**, two-sided Student's *t*-test. *P*-values are indicated.

12 weeks (Fig. 7d). Patch clamp recording demonstrated that these in vivo-converted neurons displayed functional neurophysiological properties (Extended Data Fig. 14f–i). Most notably, PTB ASO, but not control GFP ASO, rescued the 6-OHDA lesion-induced phenotype 3 months after injection, on the basis of both apomorphine-induced rotation and ipsilateral touch bias tests (Fig. 7e–g).

In summary, we report a one-step strategy to convert brain astrocytes to functional neurons. Our approach takes advantage of the genetic underpinnings of a neuronal differentiation program that is present, but latent in astrocytes. Taking advantage of the regional specificity of neuronal reprogramming, we efficiently converted midbrain astrocytes into functional DA neurons that integrate into the nigrostriatal dopamine pathway. Applying this approach to a chemically induced model of Parkinson's disease, we demonstrated partial replenishment of lost DA neurons and the restoration of striatal dopamine, leading to reversal of motor deficits. Notably, our ASO-based experiments illustrate a potentially clinically feasible approach for treatment of patients with Parkinson's disease. Eventual application of our approach to humans will need to overcome many obstacles, including age-related limits of reprogramming, understanding potential adverse effects caused by local astrocyte depletion (although we only converted only a small fraction of injury-induced astrocytes), specifically targeting regions that harbour vulnerable neurons, and detecting potential side effects due to mistargeted neurons. Each of these objectives can now be addressed experimentally to develop this promising therapeutic strategy—one that may be applicable to not only Parkinson's disease, but also other neurodegenerative disorders.

Note added in proof: While our work was under review, conceptually related results appeared elsewhere[30].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2388-4.

1. Poewe, W. et al. Parkinson disease. *Nat. Rev. Dis. Primers* **3**, 17013 (2017).
2. Barker, R. A., Götz, M. & Parmar, M. New approaches for brain repair—from rescue to reprogramming. *Nature* **557**, 329–334 (2018).
3. Sonntag, K. C. et al. Pluripotent stem cell-based therapy for Parkinson's disease: current status and future prospects. *Prog. Neurobiol.* **168**, 1–20 (2018).
4. Cohen, D. E. & Melton, D. Turning straw into gold: directing cell fate for regenerative medicine. *Nat. Rev. Genet.* **12**, 243–252 (2011).
5. Yu, X., Nagai, J. & Khakh, B. S. Improved tools to study astrocytes. *Nat. Rev. Neurosci.* **21**, 121–138 (2020).
6. Rivetti di Val Cervo, P. et al. Induction of functional dopamine neurons from human astrocytes in vitro and mouse astrocytes in a Parkinson's disease model. *Nat. Biotechnol.* **35**, 444–452 (2017).
7. Wu, Z. et al. Gene therapy conversion of striatal astrocytes into GABAergic neurons in mouse models of Huntington's disease. *Nat. Commun.* **11**, 1105 (2020).
8. Gascón, S., Masserdotti, G., Russo, G. L. & Götz, M. Direct Neuronal Reprogramming: Achievements, Hurdles, and New Roads to Success. *Cell Stem Cell* **21**, 18–34 (2017).
9. Xue, Y. et al. Direct conversion of fibroblasts to neurons by reprogramming PTB-regulated microRNA circuits. *Cell* **152**, 82–96 (2013).
10. Xue, Y. et al. Sequential regulatory loops as key gatekeepers for neuronal reprogramming in human cells. *Nat. Neurosci.* **19**, 807–815 (2016).
11. Hu, J., Qian, H., Xue, Y. & Fu, X. D. PTB/nPTB: master regulators of neuronal fate in mammals. *Biophys. Rep.* **4**, 204–214 (2018).
12. Bennett, C. F., Krainer, A. R. & Cleveland, D. W. Antisense Diseases. *Annu. Rev. Neurosci.* **42**, 385–406 (2019).
13. Guo, Z. et al. In vivo direct reprogramming of reactive glial cells into functional neurons after brain injury and in an Alzheimer's disease model. *Cell Stem Cell* **14**, 188–202 (2014).
14. Lu, T. et al. REST and stress resistance in ageing and Alzheimer's disease. *Nature* **507**, 448–454 (2014).
15. Li, Q. et al. The splicing regulator PTBP2 controls a program of embryonic splicing required for neuronal maturation. *eLife* **3**, e01201 (2014).
16. Laywell, E. D., Rakic, P., Kukekov, V. G., Holland, E. C. & Steindler, D. A. Identification of a multipotent astrocytic stem cell in the immature and adult mouse brain. *Proc. Natl Acad. Sci. USA* **97**, 13883–13888 (2000).
17. Sofroniew, M. V. Transgenic techniques for cell ablation or molecular deletion to investigate functions of astrocytes and other GFAP-expressing cell types. *Methods Mol. Biol.* **814**, 531–544 (2012).
18. Tateno, T. & Robinson, H. P. The mechanism of ethanol action on midbrain dopaminergic neuron firing: a dynamic-clamp study of the role of I(h) and GABAergic synaptic integration. *J. Neurophysiol.* **106**, 1901–1922 (2011).
19. Kimm, T., Khaliq, Z. M. & Bean, B. P. Differential regulation of action potential shape and burst-frequency firing by BK and Kv2 Channels in substantia nigra dopaminergic neurons. *J. Neurosci.* **35**, 16404–16417 (2015).
20. Boisvert, M. M., Erikson, G. A., Shokhirev, M. N. & Allen, N. J. The aging astrocyte transcriptome from multiple regions of the mouse brain. *Cell Rep.* **22**, 269–285 (2018).
21. Nott, A. et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
22. Grealish, S. et al. Human ESC-derived dopamine neurons show similar preclinical efficacy and potency to fetal neurons when grafted in a rat model of Parkinson's disease. *Cell Stem Cell* **15**, 653–665 (2014).
23. Thiele, S. L., Warre, R. & Nash, J. E. Development of a unilaterally-lesioned 6-OHDA mouse model of Parkinson's disease. *J. Vis. Exp.* **60**, 3234 (2012).
24. Beal, M. F. Parkinson's disease: a model dilemma. *Nature* **466**, S8–S10 (2010).
25. Stott, S. R. & Barker, R. A. Time course of dopamine neuron loss and glial response in the 6-OHDA striatal mouse model of Parkinson's disease. *Eur. J. Neurosci.* **39**, 1042–1056 (2014).
26. Boix, J., Padel, T. & Paul, G. A partial lesion model of Parkinson's disease in mice—characterization of a 6-OHDA-induced medial forebrain bundle lesion. *Behav. Brain Res.* **284**, 196–206 (2015).
27. Zhu, H. & Roth, B. L. DREADD: a chemogenetic GPCR signaling platform. *Int. J. Neuropsychopharmacol.* **18**, pyu007 (2015).
28. Armbruster, B. N., Li, X., Pausch, M. H., Herlitze, S. & Roth, B. L. Evolving the lock to fit the key to create a family of G protein-coupled receptors potently activated by an inert ligand. *Proc. Natl Acad. Sci. USA* **104**, 5163–5168 (2007).
29. Chen, Y. et al. Chemical control of grafted human PSC-derived neurons in a mouse model of Parkinson's disease. *Cell Stem Cell* **18**, 817–826 (2016).
30. Zhou, H. et al. Glia-to-neuron conversion by CRISPR-CasRx alleviates symptoms of neurological disease in mice. *Cell* **181**, 590-603 (2020).

Methods

Vectors and virus production

To build the lentiviral vector to express shPTB in mouse astrocytes, the target sequence 5′-GGGTGAAGATCCTGTTCAATA-3′ was shuttled into the pLKO.1-hygromycin vector (Addgene, #24150). To express shPTB in human astrocytes, the target sequence 5′-GCGTGAAGATCCTGTTCAATA-3′ was used. Viral particles were packaged in Lenti-X 293T cells (Takara bio) co-transfected with the two package plasmids: pCMV-VSV-G (Addgene, #8454) and pCMV-dR8.2 dvpr (Addgene, #8455). Viral particles were concentrated by ultracentrifugation in a Beckman XL-90 centrifuge with SW-28 rotor at 20,000 rpm for 120 min at 4 °C.

To construct AAV vectors, the same target sequence against mouse PTB was first inserted into the pTRIPZ-RFP vector (Dharmacon) between the EcoR I and Xho I sites. The segment containing RFP and shPTB was next subcloned to replace CaMP3.0 in the Asc I-digested AAV-CMV-LOX-STOP-LOX-mG-CaMP3.0 vector (Addgene, #50022). The empty vector contains only RFP subcloned into the same vector. To construct a control vector expressing non-target shRNA, the shPTB was replaced with 5′-CAACAAGATGAAGAGCACCAA-3′ to target GFP. The resulting vectors are referred to as AAV-shPTB, AAV-empty or AAV-shGFP. The AAV-hM4Di-shPTB vector was constructed by replacing RFP in AAV-shPTB with the cDNA of hM4Di, which was subcloned from pAAV-CBA-DIO-hM4Di-mCherry vector (Addgene, #81008). To express RFP and shPTB under the GFAP promoter, a segment containing floxed/off RFP and shPTB was used to replace eGFP in the AAV-GFAP-eGFP vector (Addgene, #50473) between the SalI and HindIII sites.

Viral particles of AAV2 were packaged in co-transfected HEK293T cells with the other two plasmids: pAAV-RC and pAAV-Helper (Agilent Genomics). After collection, viral particles were purified with a heparin column (GE Healthcare) and then concentrated with an Ultra-4 centrifugal filter unit (Amicon, 100,000 molecular weight cut-off). Titers of viral particles were determined by quantitative PCR to achieve $>1\times10^{12}$ particles per ml.

Synthesis of antisense oligonucleotides

ASOs were synthesized by Integrated DNA Technologies. The sequence of the target region in mouse PTB for ASO synthesis is 5′-GGGTGAAGATCCTGTTCAATA-3′, and the target sequence in Turbo GFP is 5′-CAACAAGATGAAGAGCACCAA-3′. The backbones of all ASOs contain phosphorothioate modifications. Fluorescein was attached to the 3′ end of those ASOs for fluorescence detection.

Western blot and RT–qPCR

For western blotting, cells were lysed in 1× SDS loading buffer, and after quantification, bromophenol blue was added to a final concentration of 0.1%. Protein (25–30 μg) was resolved in 10% Nupage Bis–Tris gels and probed with the primary antibodies listed in the Supplementary Table 3.

For RT–qPCR, total RNA was extracted with Trizol (Life Technology) and 10 μg ml^{-1} of glycogen was used to enhance precipitation of small RNAs. Total RNA was first treated with DNase I (Promega) followed by reverse transcription with the miScript II RT Kit (QIAGEN, 218160, for microRNA analysis) or the SuperScript III First-Strand Synthesis System (ThermoFisher, 18080051, for mRNA analysis). RT–qPCR was performed using the miScript SYBR Green PCR Kit (QIAGEN, 218073 for microRNA) or the Luna Universal qPCR Master Mix (NEB, M3003L, for mRNA) on a Step-One Plus PCR instrument (Applied Biosystems). The primers used are listed in Supplementary Table 4.

Cell culture and transdifferentiation in vitro

Mouse astrocytes were isolated from postnatal (P4–P5) pups. Cortical or midbrain tissue was dissected from whole brain and incubated with trypsin before plating onto dishes coated with poly-D-lysine (Sigma). Isolated astrocytes were cultured in DMEM (GIBCO) plus 10% fetal bovine serum (FBS) and penicillin/streptomycin (GIBCO). Dishes were carefully shaken daily to eliminate non-astrocytic cells. After reaching ~90% confluency, astrocytes were disassociated with Accutase (Innovative Cell Technologies) followed by centrifugation for 3 min at 800 rpm, and then cultured in astrocyte growth medium containing DMEM/F12 (GIBCO), 10% FBS (GIBCO), penicillin/streptomycin (GIBCO), B27 (GIBCO), 10 ng ml^{-1} epidermal growth factor (EGF, PeproTech), and 10 ng ml^{-1} fibroblast growth factor 2 (FGF2, PeproTech).

To induce transdifferentiation in vitro, mouse astrocytes were resuspended with astrocyte culture medium containing the lentivirus that targets mouse PTB, and then plated on Matrigel Matrix (Corning)-coated coverslips (12 mm). After 24 h, cells were selected with hygromycin B (100 μg ml^{-1}, Invitrogen) in fresh astrocyte culture medium for 72 h. The medium was then switched to N3/basal medium (1:1 mix of DMEM/F12 and neurobasal medium, 25 μg ml^{-1} insulin, 50 μg ml^{-1} transferring, 30 nM sodium selenite, 20 nM progesterone, 100 nM putrescine) supplemented with 0.4% B27, 2% FBS, a cocktail of 3 small molecules (1 μM ChIR99021, 10 μM SB431542 and 1mM Db-cAMP), and neurotrophic factors (brain-derived neurotrophic factor, glial cell-derived neurotrophic factor, neurotrophin 3 and ciliary neurotrophic factor, all at 10 ng ml^{-1}). The medium was half-changed every the other day. To measure synaptic currents, converted cells after 5–6 weeks were added with fresh GFP-labelled rat astrocytes, and after a further 2–3 weeks of co-culture, patch clamp recordings were performed. To test the effect of PTB ASO in vitro, mouse astrocytes were cultured in six-well plates with astrocyte growth medium. When cells reached 70%–80% confluency, PTB ASO or GFP ASO (75 pmol per well) were transfected with Lipofectamine RNAimax (ThermoFisher Scientific). Forty-eight hours after ASO treatment, cells were either collected for immunoblotting or switched to N3/basal medium for further differentiation.

Human astrocytes were purchased from Cell Applications (taken from cerebral cortex at the gestational age of 19 weeks). Cells were grown in astrocyte medium (Cell Applications) and sub-cultured until they reached ~80% confluency. For transdifferentiation in vitro, cultured human astrocytes were first disassociated with trypsin, resuspended in astrocyte medium containing the lentivirus that targets human PTB, and plated on Matrigel Matrix–coated coverslips. After 24 h, cells were selected with hygromycin B (100 μg ml^{-1}, Invitrogen) for 72 h. The medium was switched to N3/basal medium supplemented with 0.4% B27, 2% FBS and neurotrophic factors (brain-derived neurotrophic factor, glial cell-derived neurotrophic factor, neurotrophin 3 and ciliary neurotrophic factor, all at 10 ng ml^{-1}). To measure synaptic currents, converted cells after three weeks were added with fresh GFP-labelled rat astrocytes, and after a further two to three weeks of co-culture, patch clamp recordings were performed.

Other cell lines used were checked for morphology using microscopy and immunostaining with specific markers. HEK293T cells were from a common laboratory stock. Lenti-X 293T cells were purchased from Takara Bio (#632180). MEFs were isolated from E14.5 C57BL/6 mouse embryos. Mouse neurons were isolated from E17–E18 C57BL/6 mouse embryos. Human dermal fibroblasts were purchased from ATCC (PCS-201-012). Human neurons were transdifferentiated from human neuronal progenitor cells, which were a gift A. Muotri (University of California, San Diego). All cell lines tested negative for mycoplasma contamination by Hoechst staining of the cells.

RNA-seq and data analysis

Total RNA was extracted from cultured cells with the Direct-zol RNA MiniPrep kit (Zymo Research). RNA-seq was performed as previously described[31]. In brief, 2 μg of total RNA was first converted to cDNA by the superscript III first-strand synthesis kit with primer Biotin-B-T. The cDNA was purified with a PCR Clean-Up Kit (Clontech) column to remove free primer and enzyme. Terminal transferase (NEB) was applied to block the 3′ end of cDNA. Streptavidin-coated magnetic beads (Life Technologies) were used to isolate cDNAs. After RNA

degradation by sodium hydroxide, the second-strand was synthesized by random priming and then eluted from beads by heat denaturing. The cDNA was then used as template to construct RNA-seq libraries. Sequencing was run on the Hiseq 4000 system. Low-quality reads were filtered and adaptors trimmed by using the software cutadapt with parameters "-a A{10} -m 22" (ref. [32]). Cleaned reads were mapped to the pre-indexed mm10 transcriptome using the software Salmon with parameters "quant -l A –validateMappings –seqBias" (ref. [33]). Raw counts of each library were applied to the R package DEseq2 for analysis of differentially expressed genes (DEGs) with false-discovery rate < 0.05, and hierarchical clustering was performed as described[34]. The raw data from RNA-seq experiments have been deposited into NCBI under the accession number GSE142250.

Immunocytochemistry

Cultured cells grown on glass slides were fixed with 4% paraformaldehyde (PFA, Affymetrix) for 15 min at room temperature followed by permeabilization with 0.1% Triton X-100 in PBS for 15 min on ice. After washing twice with PBS, cells were blocked in PBS containing 3% BSA for 1 h at room temperature. Fixed cells were incubated with primary antibodies (listed in Supplementary Table 3) overnight at 4 °C in PBS containing 3% BSA. After washing twice with PBS, cells were incubated with secondary antibodies conjugated to Alexa Fluor 488, Alexa Fluor 546, Alexa Fluor 594 or Alexa Fluor 647 (1:500, Molecular Probes) for 1 h. DAPI (300 nM in PBS) was applied to cells for 20 min at room temperature to label nuclei. After washing three times with PBS, Fluoromount-G mounting medium was applied onto the glass slides, and images were examined and recorded under Olympus FluoView FV1000. Counting of cell numbers and percentages were all based on multiple biological replicates as detailed in specific figure legends.

For staining brain sections, mice were killed with CO_2 and immediately perfused, first with 15–20 ml saline (0.9% NaCl) and then with 15 ml 4% PFA in PBS to fix tissues. Whole brains were extracted and fixed in 4% PFA overnight at 4 °C, and then cut into 14–18 μm sections on a cryostat (Leica). Before staining, brain sections were incubated with sodium citrate buffer (10 mM sodium citrate, 0.05% Tween 20, pH 6.0) for 15 min at 95 °C for antigen retrieval. The slides were next treated with 5% normal donkey serum and 0.3% Triton X-100 in PBS for 1 h at room temperature. The remaining steps were performed as with cultured cells on coverslips.

Quantification of neuronal cell body and fibre density

Coronal sections across midbrain were sampled at intervals of 120–140 μm for immunostaining of TH and RFP. The total number (N_t) of cell types of interest was calculated by the stereological method, correcting with the Abercrombie formula[35]. The formula used is $N_t = N_s \times (S_t/S_s) \times M/(M+D)$, where N_s is the number of neurons counted, S_t is the total number of sections in the brain region, S_s is the number of sections sampled, M is the thickness of section, and D is the average diameter of counted cells, as previously described[36,37].

RFP+ and RFP+TH+ fibres were quantified using a previously published sphere method[22]. For analysing striatal fibres, three coronal sections (A/P: +1.3, +1.0 and +0.70) were selected from each brain[36]. For analysis of fibres in the nigrostriatal bundle, the coronal section closed to position Bregma −1.6 mm was selected. For each selected section, three randomly chosen areas were captured from one section of *z*-stack images at intervals of 2 μm using a 60× oil-immersion objective. A sphere (diameter: 14 μm) was then generated as a probe to measure fibre density within the whole *z*-stack. Each fibre crossing the surface of sphere was given a score. All images were analysed using Image-J 1.47v[38,39].

Electrophysiological recording

Patch clamp recordings were performed with Axopatch-1D amplifiers or Axopatch 200B amplifier (Axon Instruments) connecting to a Digidata1440A interface (Axon Instruments). Data were acquired with pClamp 10.0 or Igor 4.04 software and analysed with MatLab v2009b. For neurons in vitro converted from mouse astrocytes, small molecules were removed from medium 1 week before patch clamp recording. Both cultured mouse and human cells were first incubated with oxygenated (95% O_2 and 5% CO_2) artificial cerebrospinal fluid (150 mM NaCl, 5 mM KCl, 1 mM CaCl2, 2 mM MgCl2, 10 mM glucose, 10 mM HEPES, pH 7.4) at 37 °C for 30 min and whole-cell patch clamp was performed on selected cells.

For recording activities of in vivo converted neurons, cortical slices (300 μm) were prepared 6 or 12 weeks after injection of AAV. Brain slices were prepared with a vibratome in oxygenized (95% O_2 and 5% CO_2) dissection buffer (110.0 mM choline chloride, 25.0 mM $NaHCO_3$, 1.25 mM NaH_2PO_4, 2.5 mM KCl, 0.5 mM $CaCl_2$, 7.0 mM $MgCl_2$, 25.0 mM glucose, 11.6 mM ascorbic acid, 3.1 mM pyruvic acid) at 4 °C followed by incubation in oxygenated artificial cerebrospinal fluid (aCSF) (124 mM NaCl, 3 mM KCl, 1.2 mM NaH_2PO_4, 26 mM $NaHCO_3$, 2.4 mM $CaCl_2$, 1.3 mM $MgSO_4$, 10 mM dextrose and 5 mM HEPES; pH 7.4) at room temperature for 1 h before experiments.

Patch pipettes (5–8 MΩ) solution contained 150 mM KCl, 5 mM NaCl, 1 mM $MgCl_2$, 2 mM EGTA–Na, and 10 mM Hepes pH 7.2. For voltage-clamp experiments, the membrane potential was typically held at −75 mV. The following concentrations of channel blockers were used: PiTX: 50 μM; NBQX: 20 μM; APV: 50 μM. All of these blockers were applied to the bath following dilution into the external solution from concentrated stock solutions. All experiments were performed at room temperature.

Transgenic mice

The *Gfap-cre* transgenic mouse (B6.Cg-Tg(*Gfap-cre*)77.6Mvs/2J) was used in AAV-shPTB induced in vivo reprogramming experiments. The Slc6a3-Cre transgenic mouse (B6.SJL-Slc6a3tm1.1(cre)Bkmn/J) was used for chemogenetic experiments. For testing the effect of ASOs in vivo, the *Gfap-creER*^{TM} mouse (B6.Cg-Tg(*Gfap-cre/ERT2*)505Fmv/J) was crossed with the Rosa-tdTomato mouse (B6.Cg-Gt(ROSA) 26Sortm14(CAG-tdTomato)Hze/J). Offspring of these double *Gfap-CreER*^{TM};*Rosa-tdTomato* transgenic mice at age P30–P40 days were injected with tamoxifen (dissolved in corn oil at a concentration of 20 mg ml^{-1}) via intraperitoneal injection once every 24 h for a total of 5 consecutive days. The dose of each injection was 75 mg kg^{-1}. Two weeks after tamoxifen administration, PTB ASO or control ASO was injected into substantia nigra of those mice to investigate ASO-induced in vivo reprogramming.

All transgenic mice were purchased from The Jackson Laboratory. All procedures were conducted in accordance with the guide of The University of California San Diego Institutional Animal Care and Use Committee (protocol no. S99116). Both male and female mice were used and randomly grouped in this study. No pre-tests were performed to determine sample sizes. Most studies used mice aged P30–P40. As indicated in Fig. 6d, e, mice at 1 year of age were also tested for AAV-shPTB-mediated reprogramming and behavioural tests.

Ipsilateral lesion with 6-OHDA and stereotaxic injections

Adult WT and *Gfap-cre* mice at P30–P40 were used to perform surgery to induce lesion. Animals were anaesthetized with a mix of ketamine (80-100 mg/kg) and xylazine (8-10 mg/kg) and then placed in a stereotaxic mouse frame. Before injecting 6-hydroxydopamine (6-OHDA, Sigma), mice were treated with a mix of desipramine (25 mg/kg) and pargyline (5 mg/kg). 6-OHDA (3.6 μg per mouse) was dissolved in 0.02% ice-cold ascorbate/saline solution at a concentration of 15 mg/ml and used within 3 h. The toxic solution was injected into the medial forebrain bundle at the following coordinates (relative to bregma): anterior–posterior (A/P) = -1.2mm; medio-lateral (M/L) = 1.3mm and dorso-ventral (D/V) = 4.75mm (from the dura). Injection was applied in a 5 μl Hamilton syringe with a 33G needle at the speed of 0.1 μl/min. The needle was slowly removed 3 min after injection. Cleaning and suturing of the wound were performed after lesion.

AAVs or ASOs were injected into substantia nigra ~30 days after 6-OHDA induced lesion. Four microlitres of AAV or 2 µl of ASO (1 µg µl^{-1}) was injected into lesioned substantia nigra at the following coordinates A/P: −3.0 mm; M/L: 1.2 mm and D/V: 4.5 mm. Injections were made using the same syringe and needle, at a rate of 0.5 µl min^{-1}. The needle was slowly removed 3 min after injection. For injecting AAV in striatum and visual cortex, the following coordinates were employed: A/P: +1.2 mm; M/L: 2.0 mm; D/V 3.0 mm (for striatum), and A/P: −4.5 mm; M/L: 2.7 mm; D/V: 0.35 mm (visual cortex).

Retrograde tracing

For retrograde tracing of the nigrostriatal pathway, *Gfap-cre* mice with or without 6-OHDA induced lesion were first injected with AAV-shPTB. 1 or 3 months after AAV delivery, green Retrobeads IX (Lumafluor) were unilaterally injected at two sites into the striatum on the same side of AAV injection, using following two coordinates: A/*P*: + 0.5mm, M/L: 2.0 mm; D/V: 3.0 mm; and A/P: +1.2 mm; M/L: 2.0 mm; D/V: 3.0 mm. Approximately 2 µl of beads was injected. After 24 h, animals were killed and immediately perfused. Their brains were fixed with 4% PFA for sectioning and immunostaining.

Measurement of striatal dopamine

Dopamine levels in mouse striatum were measured by reverse-phase HPLC. The HPLC analysis was performed on an Agilent 1260 Infinity HPLC system with an Agilent Zorbax SB-C18 semi-prep column (ID 9.4 × 250 mm, 5 µm, 80Å) using a water/methanol gradient containing 0.1% formic acid. Each substance was characterized by retention time and 260 nm absorbance under a variable wavelength detector as previously described[40,41]. Striatal samples were directly prepared from brain tissue. In brief, striatal dissection was carried out immediately after euthanization. After homogenization in 200 µl of 0.1 M hydrochloric acid with a Squisher homogenizer, the sample was centrifuged (12,000*g*, 10 min, 4 °C). The resulting supernatant was filtered by a 0.2 µm Nanosep MF centrifugal device and then analysed by HPLC[41,42]. Investigators were blinded to group identity for measurements of striatal dopamine.

Amperometric dopamine recording

The amperometric recording of dopamine release in vivo was conducted, as described previously[43,44]. Anaesthetized mice were fixed on a stereotaxic instrument (Narishige). Body temperature was monitored and maintained at 37 °C using a heating pad (KEL-2000). A bipolar stimulating electrode was implanted in the medial forebrain bundle (MFB: 2.1 mm AP, 1.1 mm ML, 4.0–5.0 mm DV). The recording carbon fibre electrode (7 µm diameter, 400 µm long) was implanted in the caudate putamen of dorsal striatum (CPu: 1.1 mm AP, 1.7 mm ML, 3.4 mm DV). An Ag–AgCl reference electrode was placed in the contralateral cortex. Electric stimulation was generated using an isolator (A395, WPI) as a train of biphasic square-wave pulses (0.6 mA, 1 ms duration, 36 pulses, 80 Hz). The carbon fibre electrode was maintained at 780 mV to oxidize the substance. The amperometric signal was amplified by a patch-clamp amplifier (PC2C, INBIO), low pass-filtered at 50 Hz and recorded by MBA-1 DA/AD unit v4.07 (INBIO). Investigators were blinded to group identity for measurements of dopamine release.

Amperometric recordings of dopamine release on dorsal striatum slices were conducted, as described previously[45,46]. Anaesthetized mice were transcardially perfused with ~20 ml ice-cold recording aCSF containing 110 mM $C_5H_{14}NClO$, 2.5 mM KCl, 0.5 mM $CaCl_2$, 7 mM $MgCl_2$, 1.3 mM NaH_2PO_4, 25 mM $NaCO_3$, 25 glucose (saturated with 95% O_2 and 5% CO_2). The brain was rapidly removed and cut into 300-µm horizontal slices on a vibratome (Leica VT1000s) containing ice-cold sectioning solution. Slices containing striatum were allowed to recover for 30 min in recording aCSF: 125 mM NaCl, 2.5 mM KCl, 2 mM $CaCl_2$, 1.3 mM $MgCl_2$, 1.3 mM NaH_2PO_4, 25 mM $NaCO_3$, 10 mM glucose (saturated with 95% O_2 and 5% CO_2) at 37 °C, and then kept at room temperature for recording. Carbon fibre electrodes (7 µm diameter, 200 µm long) holding at 780 mV were used to measure dopamine release in striatum. The exposed carbon fibre electrode tip was completely inserted into the subsurface of the striatal slice at an angle of ~30°. Single electrical field stimulation pulses (0.2 ms, 0.6 mA) were delivered through a bipolar platinum electrode (150 µm in diameter) and generated by a Grass S88K stimulator (Astro-Med). The amperometric current (I_{amp}) was low-pass filtered at 100 Hz and digitized at 3.13 kHz. Off-line analysis was performed using Igor software (WaveMetrix). Amperometric recording in cultured cells was conducted as previously described[47]. Reprogrammed neurons were pre-treated with 100 µM 3,4-dihydroxyphenylalanine (L-DOPA) for 30 min for signal enhancement. During recording, carbon fibre electrodes (WPI, CF30-50) were held at +750 mV to measure dopamine release. For baseline recording, cells were kept in normal aCSF (150 mM NaCl, 5 mM KCl, 1 mM $CaCl_2$, 2 mM $MgCl_2$, 10 mM glucose, 10 mM HEPES, pH 7.4). The solution was then switched to a high potassium aCSF (130 mM NaCl, 25 mM KCl, 1 mM $CaCl_2$, 2 mM $MgCl_2$, 10 mM glucose, 10 mM HEPES, pH 7.4) to induce the release of dopamine. No spike-like events were detected when the electrode was held at −750 mV under the same conditions[47].

Behavioural testing

All behavioural tests were carried out 21–28 days after 6-OHDA induced lesion or 2, 3 and 5 months after the delivery of AAVs or ASOs. The investigators were not blinded to allocation during experiments and outcome assessment. For the rotation test, apomorphine-induced rotations in mice were recorded after intraperitoneal injection of apomorphine (Sigma, 0.5 mg kg^{-1}) under a live video system. Mice were injected with apomorphine (0.5 mg kg^{-1}) on two separate days before performing the rotation test (for example, if the test was to be performed on Friday, the mouse would be first injected on Monday and Wednesday), which aimed to prevent a 'wind-up' effect that could obscure the final results. Rotation was measured 5 min following the injection for 10 min, as previously described[48,49] and only full-body turns were counted. Rotations induced by D-amphetamine (Sigma, 5 mg kg^{-1}) were determined in the same system[50,51]. Data were expressed as net contralateral or ipsilateral turns per min.

To perform the cylinder test, mice were individually placed into a glass cylinder (diameter 19 cm, height 25 cm), with mirrors placed behind for a full view of all touches, as described[26,48]. Mice were recorded under a live video system and no habituation of the mice to the cylinder was performed before recording. A frame-by-frame video player (KMPlayer v.4.0.7.1) was used for scoring. Only wall touches independently with the ipsilateral or the contralateral forelimb were counted. Simultaneous wall touches (touched executed with both paws at the same time) were not included in the analysis. Data are expressed as a percentage of ipsilateral touches in total touches.

For chemogenetic experimenst, cylinder tests were carried out 21–28 days after 6-OHDA induced lesion and 2 months after the delivery of AAV-hM4Di-shPTB. In the later test, each animal was first injected with saline to record the baseline of recovery. Subsequent recording was performed 40 min after intraperitoneal injection of CNO (Biomol International, 4 mg kg^{-1}) or 72 h after metabolism of the drug[29].

Data analysis and statistics

The numbers (*n*) of biological replicates or mice are indicated in individual figure legends. The experiments were not randomized and no statistical methods were used to predetermine sample size. Experimental variations in each graph were represented as mean ± s.e.m. All measurements were performed on independent samples. Independent *t*-test, one-way ANOVA and repeat-measurement ANOVA were employed for statistical analysis, as indicated in individual figure legends. For multiple comparisons, combining ANOVA, post hoc Tukey test was applied. Assumptions of normal data distribution and homoscedasticity were adopted in *t*-test and one-way ANOVA. All statistical tests were

two-sided. For Fig. 1b and Extended Data Fig. 1b, the original data were transformed to logarithm with base 10 for one-way ANOVA to fulfill the requirement of homoscedasticity. To estimate the effect size, Cohen's *d* for *t*-test and eta-squared (η^2) for one-way ANOVA were calculated as previously described[52,53]. Statistical report for all figure panels is summarized in Supplementary Table 5.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

RNA-seq data have been deposited in the NCBI Gene Expression Omnibus under accession number GSE142250. Independently generated data are available upon request. Methods have been converted into stepwise protocols and deposited in Protocol Exchange (doi:10.21203/rs.3.pex-902/v1). Repeats of individual experiments are summarized in Supplementary Table 2, which has been independently verified. All data generated or analysed in this study are included in this published article (and its Supplementary Information files).

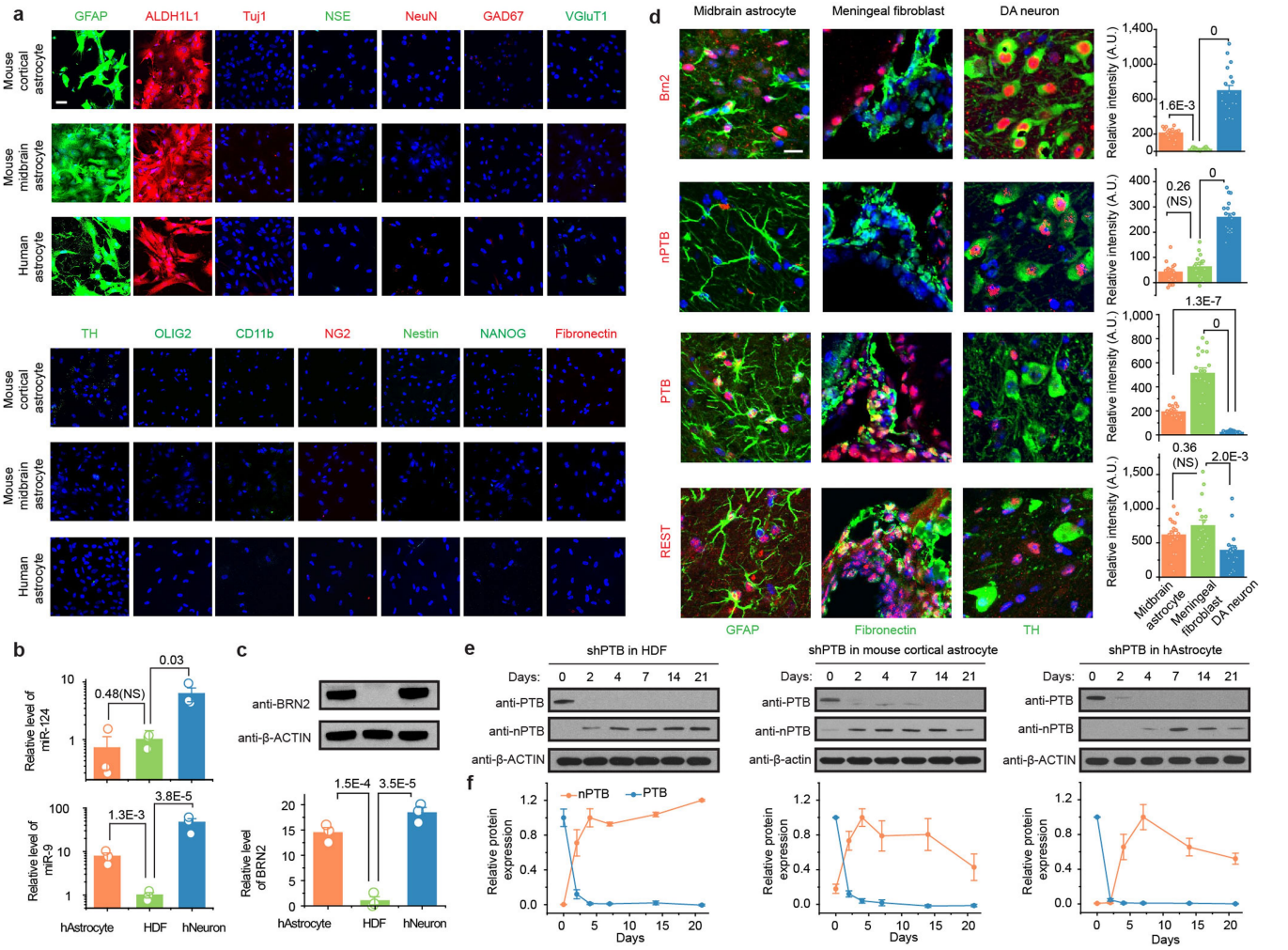
31. Ouyang, H. et al. WNT7A and PAX6 define corneal epithelium homeostasis and pathogenesis. *Nature* **511**, 358–361 (2014).
32. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
33. Srivastava, A., Malik, L., Smith, T., Sudbery, I. & Patro, R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol*. **20**, 65 (2019).
34. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15**, 550 (2014).
35. Abercrombie, M. Estimation of nuclear population from microtome sections. *Anat. Rec*. **94**, 239–247 (1946).
36. Falk, T. et al. Vascular endothelial growth factor-B is neuroprotective in an in vivo rat model of Parkinson's disease. *Neurosci. Lett*. **496**, 43–47 (2011).
37. Baker, H., Joh, T. H. & Reis, D. J. Genetic control of number of midbrain dopaminergic neurons in inbred strains of mice: relationship to size and neuronal density of the striatum. *Proc. Natl Acad. Sci. USA* **77**, 4369–4373 (1980).
38. Kordower, J. H. et al. Neurodegeneration prevented by lentiviral vector delivery of GDNF in primate models of Parkinson's disease. *Science* **290**, 767–773 (2000).
39. Bahat-Stroomza, M. et al. Induction of adult human bone marrow mesenchymal stromal cells into functional astrocyte-like cells: potential for restorative treatment in Parkinson's disease. *J. Mol. Neurosci*. **39**, 199–210 (2009).
40. Liu, G., Chen, J. & Ma, Y. Simultaneous determination of catecholamines and polyamines in PC-12 cell extracts by micellar electrokinetic capillary chromatography with ultraviolet absorbance detection. *J. Chromatogr. B* **805**, 281–288 (2004).
41. De Benedetto, G. E. et al. A rapid and simple method for the determination of 3,4-dihydroxyphenylacetic acid, norepinephrine, dopamine, and serotonin in mouse brain homogenate by HPLC with fluorimetric detection. *J. Pharm. Biomed. Anal*. **98**, 266–270 (2014).
42. Tareke, E., Bowyer, J. F. & Doerge, D. R. Quantification of rat brain neurotransmitters and metabolites using liquid chromatography/electrospray tandem mass spectrometry and comparison with liquid chromatography/electrochemical detection. *Rapid Commun. Mass Sp*. **21**, 3898–3904 (2007).
43. Wang, S. R. et al. Role of vesicle pools in action potential pattern-dependent dopamine overflow in rat striatum in vivo. *J. Neurochem*. **119**, 342–353 (2011).
44. Xu, H. et al. Striatal dopamine release in a schizophrenia mouse model measured by electrochemical amperometry in vivo. *Analyst* **140**, 3840–3845 (2015).
45. Wang, C. et al. Synaptotagmin-11 is a critical mediator of parkin-linked neurotoxicity and Parkinson's disease-like pathology. *Nat. Commun*. **9**, 81 (2018).
46. Wang, L. et al. Modulation of dopamine release in the striatum by physiologically relevant levels of nicotine. *Nat. Commun*. **5**, 3925 (2014).
47. Caiazzo, M. et al. Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. *Nature* **476**, 224–227 (2011).
48. Grealish, S., Mattsson, B., Draxler, P. & Björklund, A. Characterisation of behavioural and neurodegenerative changes induced by intranigral 6-hydroxydopamine lesions in a mouse model of Parkinson's disease. *Eur. J. Neurosci*. **31**, 2266–2278 (2010).
49. Piallat, B., Benazzouz, A. & Benabid, A. L. Subthalamic nucleus lesion in rats prevents dopaminergic nigral neuron degeneration after striatal 6-OHDA injection: behavioural and immunohistochemical studies. *Eur. J. Neurosci*. **8**, 1408–1414 (1996).
50. Dunnett, S. B., Björklund, A., Stenevi, U. & Iversen, S. D. Behavioural recovery following transplantation of substantia nigra in rats subjected to 6-OHDA lesions of the nigrostriatal pathway. *I. Unilateral lesions. Brain Res*. **215**, 147–161 (1981).
51. Iancu, R., Mohapel, P., Brundin, P. & Paul, G. Behavioral characterization of a unilateral 6-OHDA-lesion model of Parkinson's disease in mice. *Behav. Brain Res*. **162**, 1–10 (2005).
52. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, 1988).
53. Cohen, J. Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educ. Psychol. Meas*. **33**, 107–112 (1973).

Acknowledgements We thank members of the Fu laboratory for cooperation, reagent sharing and insightful discussion during the course of this investigation and A. Muotri for the gift of the human embryonic stem cell-derived neural progenitors. D.W.C. received a salary from the Ludwig Institute for Cancer Research and is a Nomis Foundation Distinguished Scientist. Z.Z. and X.K. were supported by NSFC grants (31930061, 31761133016, 21790394 and 81974203). W.C.M. and X.-D.F. were supported by a grant from the Larry Hillblom Foundation (2019-A-006-NET). This work was supported by NIH grants (GM049369 and GM052872) to X.-D.F.

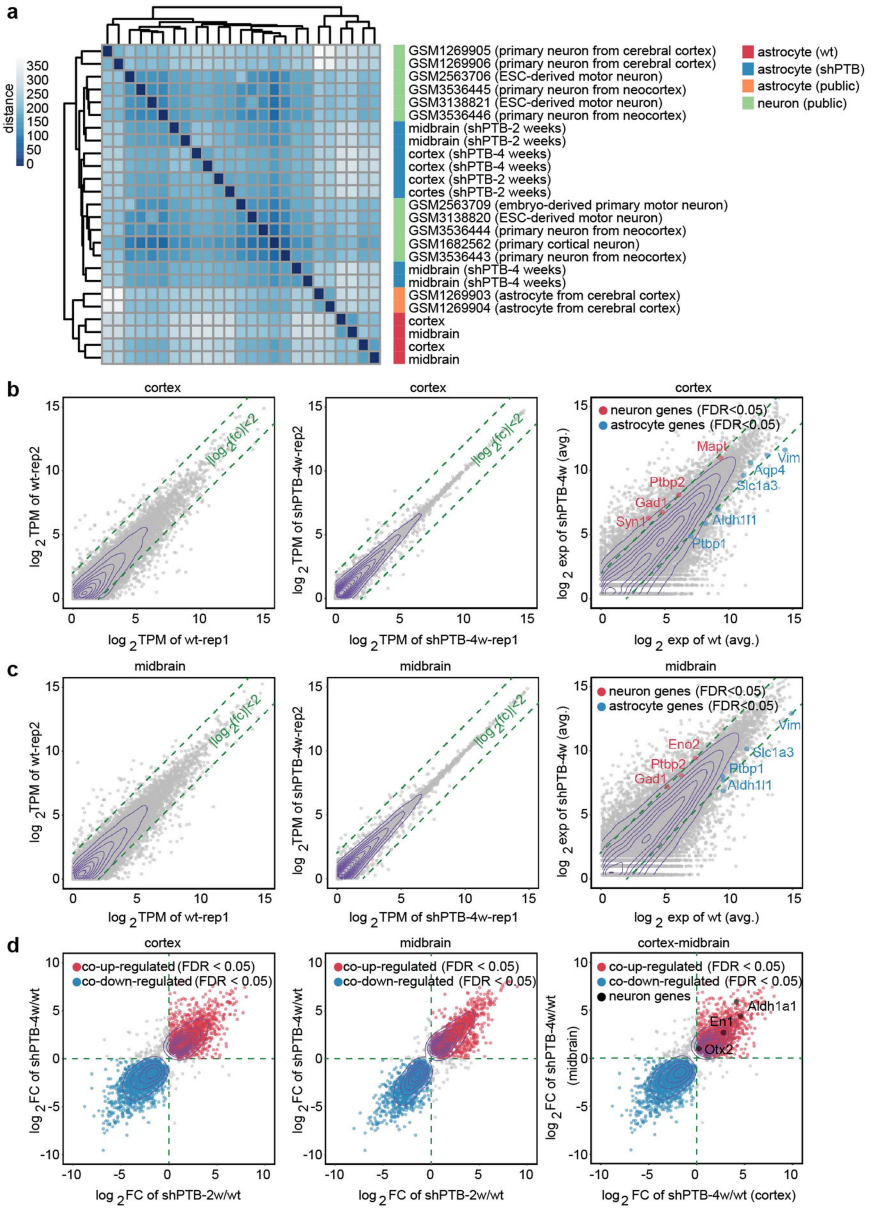
Author contributions H.Q. and X.-D.F. designed the study. H.Q. performed astrocyte isolation, stereotaxic injection, immunocytochemistry, electrophysiological measurements and behaviour tests. J.H., Y.X. and F.M. contributed to AAV vector construction, immunoblotting and immunocytochemistry and independently characterized astrocyte conversion in vitro. F.M. also contributed to the performance of behaviour tests. Z.L. and F.M. performed RNA-seq and data analysis. H.Q., X.Z., D.Z. and N.K.D. measured striatal dopamine levels. X.K. and Z.Z. recorded activity-induced dopamine release in live animals and on brain slices. S.F.D. supervised ASO design and testing. W.C.M. contributed to analysis and interpretation of neurological data. D.W.C. oversaw biochemical and immunocytochemistry experiments. R.M. independently showed that ASO-mediated suppression of PTB generated new neurons in wild-type mice. R.M. also checked all raw data and verified biological repeats of individual experiments. H.Q., S.F.D., W.C.M., D.W.C. and X.-D.F. wrote the paper.

Competing interests X.-D.F. is a founder of CurePharma. The University of California, San Diego has filed a patent application on neuronal reprogramming induced by inactivating PTB by any means for treatment of neurological disorders.

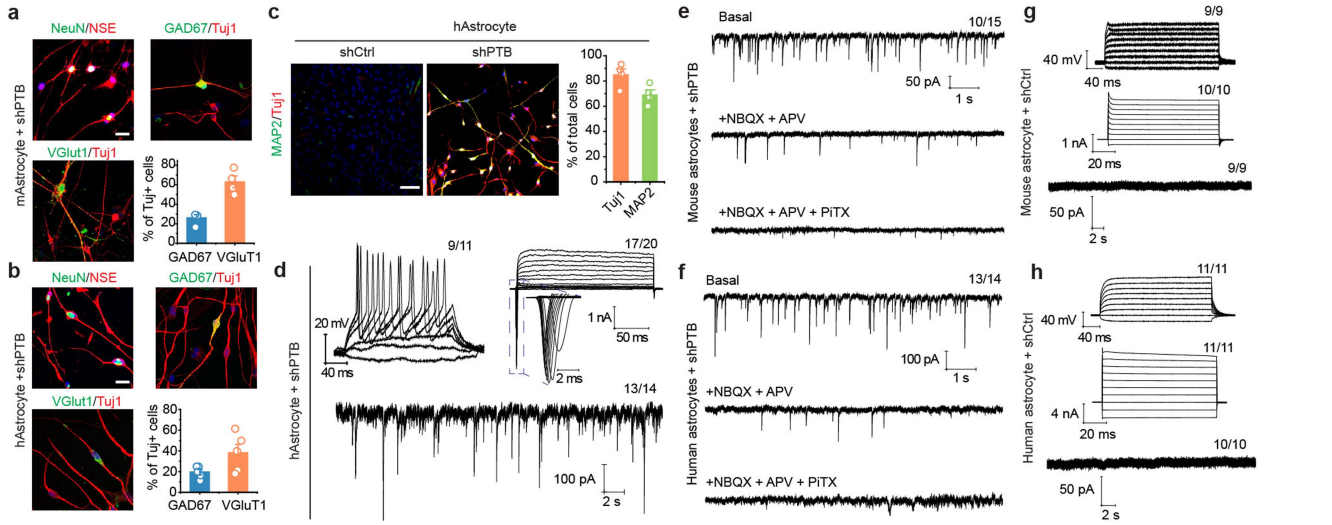
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2388-4.
Correspondence and requests for materials should be addressed to X.-D.F.
Peer review information *Nature* thanks Ernest Arenas, Anders Bjorklund, Aaron D. Gitler, Malin Parmar and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



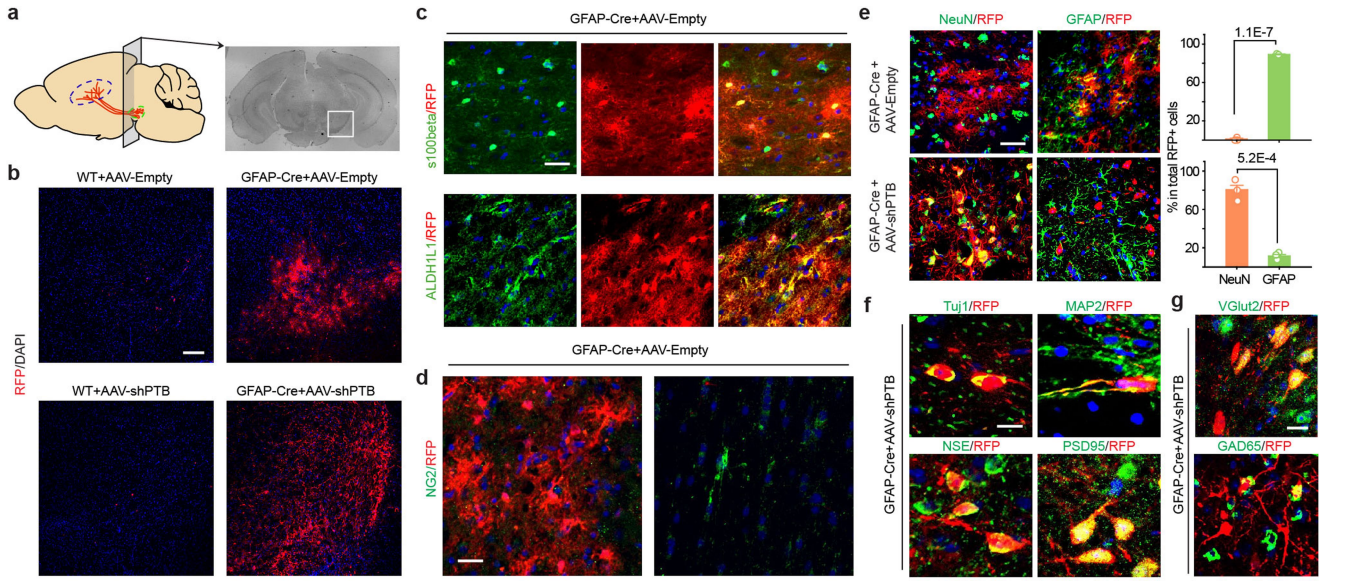
Extended Data Fig. 1 | Characterization and functional analysis of astrocytes from mice and humans. a, Relative purity of mouse and human astrocytes. Astrocytes isolated from mouse cortex and midbrain or obtained from human embryonic brain (gestational age 19 weeks) were probed with a panel of markers for neurons and common non-neuronal cell types in the central nervous system, including those for astrocytes: GFAP (green) and ALDH1L1 (red); for neurons: TUJ1, NSE, NeuN, GAD67, VGlut1 and TH; for oligodendrocytes: OLIG2; for microglia: CD11b; for NG2 cells: NG2; for neural progenitors: nestin; for pluripotent stem cells: NANOG; and for fibroblasts: fibronectin. Scale bar, 30 µm. These results demonstrated that isolated astrocytes are largely free of neurons and common non-neuronal cells. The experiment was independently repeated twice with similar results. **b**, **c**, Levels of key components in the regulatory loops controlled by PTB and nPTB in mouse midbrain. Levels of miR-124 (**b**, top) and miR-9 (**b**, bottom) were quantified by RT–qPCR in human astrocytes (hAstrocytes), human dermal fibroblasts (HDFs), and human neurons (hNeurons) differentiated from human neuronal progenitor cells. Data were normalized against U6 snRNA and the levels in human dermal fibroblasts were set to 1 for comparative analysis. Levels of BRN2 were determined by western blotting and normalized against β-actin (**c**). Results show low miR-124, but high miR-9 and BRN2 in human astrocytes, suggesting that the PTB-regulated loop is inactive and components of the nPTB-regulated loop are active in human astrocytes. **d**, Levels of PTB, nPTB, BRN2 and REST in mouse midbrain. Cell types in mouse midbrain were marked by GFAP for astrocytes, TH for DA neurons, and fibronectin for adjacent meningeal fibroblasts and double-stained for BRN2, PTB, nPTB and REST. Scale bar, 20 µm. Relative immunofluorescence intensities in different cell types were quantified (right). $n = 3$ mice with a total of 54 cells counted in each. Note that REST is decreased, but not eliminated, in endogenous DA neurons, which is in agreement with the documented requirement for REST for viability of mature neurons. **e**, **f**, Dynamic nPTB expression in response to PTB knockdown. nPTB expression was monitored by western blotting after PTB knockdown in human dermal fibroblasts (**e**, left), mouse cortical astrocytes (**e**, middle) and human astrocytes (**e**, right). **f**, Data from 3 biological repeats were quantified. Results show that nPTB remains stably expressed in human dermal fibroblasts, but undergoes transient expression in astrocytes from both mice and humans. In **b**–**d**, ANOVA with post hoc Tukey test; mean ± s.e.m. ($n = 3$ biological repeats). *P*-values are indicated. All except those pairwise comparisons indicated as NS (not significant) in panels **b** and **d** are considered statistically significant.



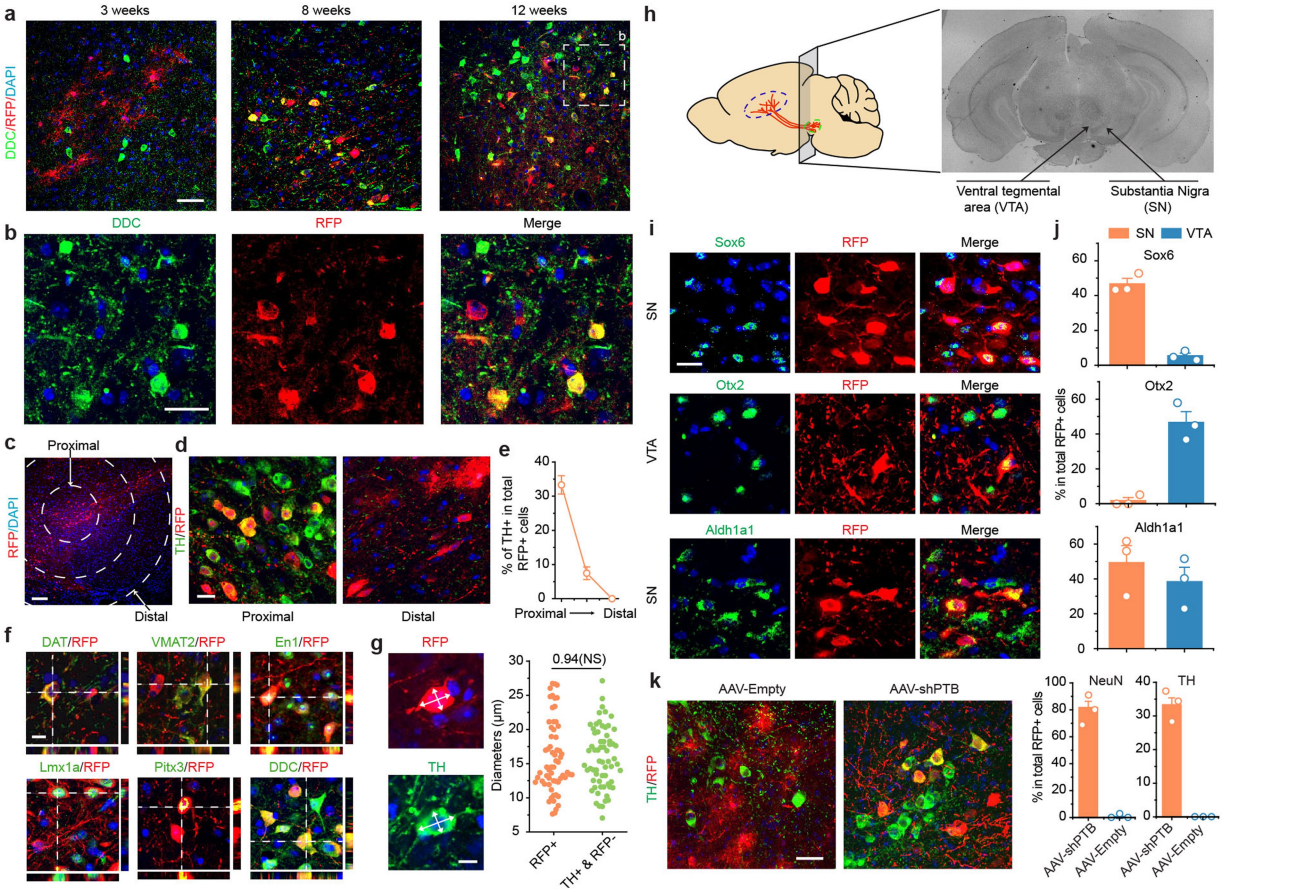
Extended Data Fig. 2 | Global evidence for programmed switch of gene expression from astrocytes to neurons in response to PTB depletion. **a**, Clustering analysis. RNA-seq data (available under GSE142250) were generated on independent isolates of astrocytes from mouse cortex or midbrain before and after conversion to neurons by depleting PTB for 2 or 4 weeks. By clustering analysis, the global gene expression profiles were compared with the public datasets for astrocytes or neurons as indicated by the colour key and the data sources on the right. The selection of these public data for comparison was based on astrocytes without further culture and on neurons directly isolated from mouse brain or differentiated from embryonic stem cells (ESCs). **b**, **c**, Comparison of gene expression profiles between independent libraries prepared from mouse cortical (**b**) or midbrain (**c**) astrocytes before and after PTB depletion for 2 or 4 weeks. Selective astrocyte-specific (blue) and neuron-specific (red) genes are highlighted. Results show a degree of heterogeneity between independent isolates of astrocytes, but notably, their converted neurons became more homogeneous. **d**, Comparison between induced gene expression upon PTB depletion in cortical versus midbrain astrocytes. Several commonly induced DA neuron-specific genes (that is, *Otx2*, *En1* and *Aldh1a1*) are highlighted when comparing between neurons derived from cortical versus midbrain astrocytes (right). Significantly induced DA neuron-associated genes are listed in Supplementary Table 1. Note that most genes are enriched, but not uniquely expressed, in DA neurons (thus, they are not specific markers for DA neurons), as evidenced by their induction to different degrees in shPTB-treated cortical astrocytes.



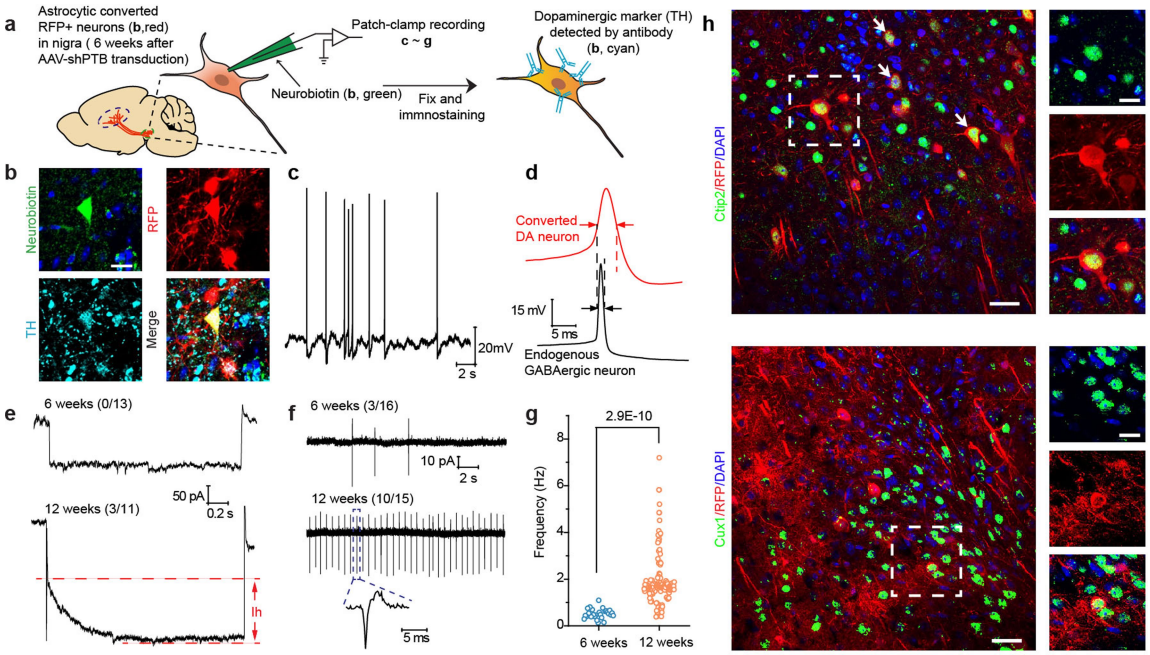
Extended Data Fig. 3 | Characterization of converted neurons from mouse and human astrocytes. a, **b**, Conversion of mouse and human astrocytes to neurons. Cells were immunostained with the indicated markers after conversion from mouse cortical astrocytes (**a**) or human astrocytes (**b**). Converted glutamatergic (marked by VGlut1) and GABAergic (marked by GAD67) neurons constituted approximately 90% and 80% of total TUJ1-marked neurons from mouse and human astrocytes, respectively. Data were based on 4 (**a**) or 5 (**b**) biological repeats and represented as mean ± s.e.m. Scale bars, 30 μm (**a**); 40 μm (**b**). **c**, **d**, Efficient conversion from human astrocytes to neurons. Converted neurons were characterized by immunostaining with TUJ1 and MAP2 (**c**). Scale bar, 80 μm. $n = 4$ biological repeats. **d**, These neurons are functional as indicated by repetitive action potentials (top left), large currents of voltage-dependent sodium and potassium channels (top right) and spontaneous postsynaptic currents after co-culture with rat astrocytes (bottom). Indicated in each panel is the number of cells that showed the recorded activity versus the number of cells examined. **e**–**h**, Electrophysiological characterization of neurons converted from mouse (**e**) and human (**f**) astrocytes, showing spontaneous excitatory and inhibitory postsynaptic currents that could be sequentially blocked with the inhibitors against the excitatory (NBQX and APV) and inhibitory (PiTX) receptors, indicative of their secretion of glutamine and GABA neurotransmitters. **g**, **h**, Control shRNA (shCtrl)-treated mouse (**g**) and human (**h**) astrocytes failed to show action potentials (top), currents of voltage-dependent channels (middle) or postsynaptic events (bottom). The number of cells that showed the recorded activity versus the total number of cells examined is indicated on the top right of each panel.



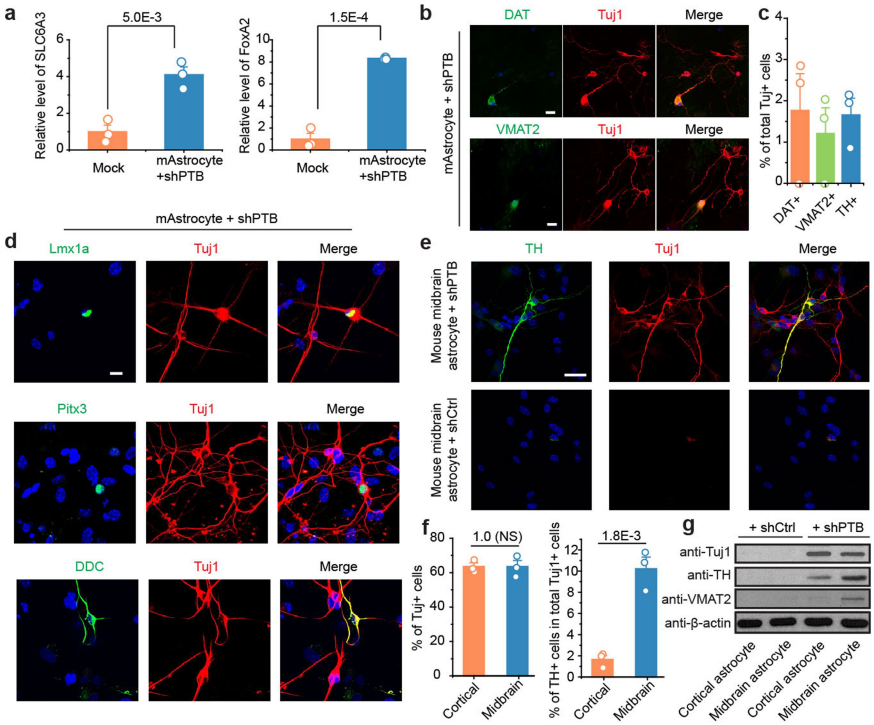
Extended Data Fig. 4 | Cre-dependent expression of RFP in injected mouse midbrain. a, Schematic of the substantia nigral region (white box) for AAV injection and immunochemical analysis. **b**, Cre-dependent RFP expression. RFP+ cells were not detected in midbrain of wild-type mice injected with either AAV-empty or AAV-shPTB (left). In comparison, both viruses generated abundant RFP signals in *Gfap-cre* transgenic mice. Scale bar, 150 μm. **c**, **d**, Co-staining of RFP+ cells with the astrocyte markers S100b and ALDH1L1 10 weeks after injecting AAV-empty (**c**), indicating that most RFP+ cells in AAV-empty-transduced midbrain were astrocytes. Scale bar, 25 μm. **d**, No RFP expression was detectable in NG2-labelled cells. Scale bar, 15 μm. Experiments in **b**–**d** were independently repeated three times with similar results. **e**, Reprogramming-dependent conversion from astrocytes to neurons. Immunostaining with the astrocyte marker GFAP and the pan-neuronal marker NeuN was performed 10 weeks after injection of AAV-empty or AAV-shPTB in the midbrain. Scale bar, 30 μm. Quantified results show that cells transduced with AAV-empty were all GFAP+ astrocytes, whereas cells transduced with AAV-shPTB were mostly NeuN+ neurons. Quantified data were based on three mice as shown on the right. Two-sided Student's *t*-test. Data are mean ± s.e.m. **f**, **g**, Further characterization of AAV-shPTB-induced neurons in midbrain with additional neuronal markers, including pan-neuronal specific markers TUJ1, MAP2, NSE and PSD95 (**f**; scale bar, 10 μm) and specific markers for glutamatergic (VGlut2) and GABAergic (GAD65) neurons (**g**; scale bar, 20 μm).



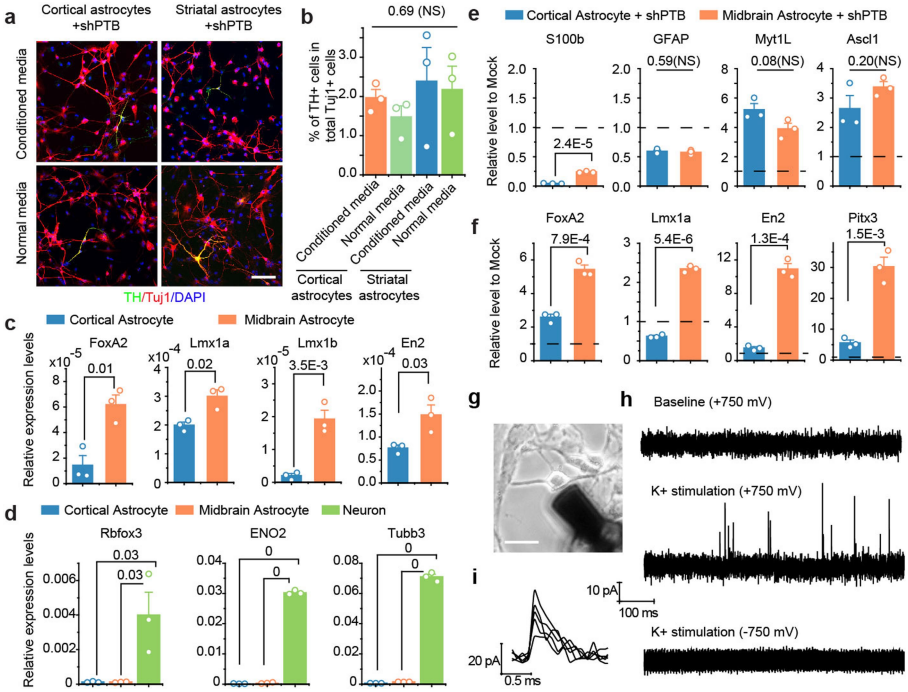
Extended Data Fig. 5 | Progressive conversion of AAV-shPTB treated astrocytes to DA neurons within the dopamine domain. a, **b**, Time-dependent appearance of RFP+DDC+ DA neurons. AAV-shPTB-transduced midbrain was characterized for time-dependent appearance of DA neurons with the DA neuron marker DDC (**a**; scale bar, 50 μm). Few initial RFP+ cells were co-stained with DDC 3 weeks after AAV-shPTB transduction, and the fraction of RFP+DDC+ cells progressively increased 8 and 12 weeks after AAV-shPTB injection. Images from substantia nigra 12 weeks after AAV-shPTB transduction are enlarged to highlight RFP+DDC+ neurons (**b**; scale bar, 25 μm). **c**–**e**, Conversion of midbrain astroyctes to DA neurons within the dopamine domain. AAV-shPTB-induced neuronal reprogramming was determined relative to the site of injection. **c**, A low-magnification view of a substantia nigra section. Circles mark brain areas with progressively larger diameters from the centre of the injection site. Scale bar, 100 μm. **d**, Enlarged views show the representative proximal and distal sites from the injection site 12 weeks after AAV-shPTB transduction, positively stained for TH (green) over RFP-labelled cells. Scale bar, 10 μm. Note the presence of RFP+TH+ cells in the proximal site, but only RFP+TH− cells in the distal site. **e**, The percentages of TH+ cells among total RFP+ cells in the three different areas defined in (**c**) were quantified based on 3 mice with at least 100 cells counted in each. Data are mean ± s.e.m. These data show the generation of TH+ neurons within the dopamine domain of midbrain. **f**, **g**, Further characterization of converted DA neurons with additional DA neuron-specific markers DAT, VMAT2, EN1, LMX1A, PITX3 and DDC, all showing positive signals (**f**). RFP+TH+ cell bodies are highlighted by orthogonal views of *z*-stacked images, attached on right and bottom of the main image (**f**; scale bar, 10 μm). Cell body diameters were compared between newly converted RFP+TH+ neurons and endogenous RFP−TH+ DA neurons (**g**, left; scale bar, 5 μm). The size distribution of both populations of neurons shown on the right suggests that converted TH+ cells have a similar cell size to endogenous TH+RFP− DA neurons (**g**, right). Quantification based on 62 RFP+ cells and 64 RFP−TH+ cells from 3 mice. Two-sided Student's *t*-test. **h**, Schematic depiction for further analysis of converted neurons in substantia nigra and ventral tegmental area. **i**, **j**, Representative immunostaining of SOX6, OTX2 and ALDH1A1, showing that SOX6-marked RFP+ cells were confined to the substantia nigra, whereas OTX2-marked RFP+ cells were in the ventral tegmental area; the DA neuron marker ALDH1A1 was detected in both substantia nigra and ventral tegmental area (**i**; scale bar, 25 μm). **j**, Quantification based on 3 mice with at least 100 cells counted. Data are mean ± s.e.m. Results further support the generation of different subtypes of DA neurons. **k**, Minimal leaky Cre expression in endogenous DA neurons in midbrain. As *Gfap-cre* is known to show a degree of leaky expression in neurons, raising a concern that AAV-shPTB might infect some endogenous DA neurons, mice treated with AAV-empty (which expresses RFP but not shPTB) were examined carefully. Scale bar, 30 μm. Compared with AAV-shPTB treated mice, few RFP+ cells stained positively for either NeuN or TH in the midbrain of mice transduced with AAV-empty, as quantified on the right, based on 3 mice with at least 100 cells counted in each. Data are mean ± s.e.m. Results show little, if any, leaky Cre expression in endogenous DA neurons and in midbrain regions of mice at the age (two months old) used in our studies.



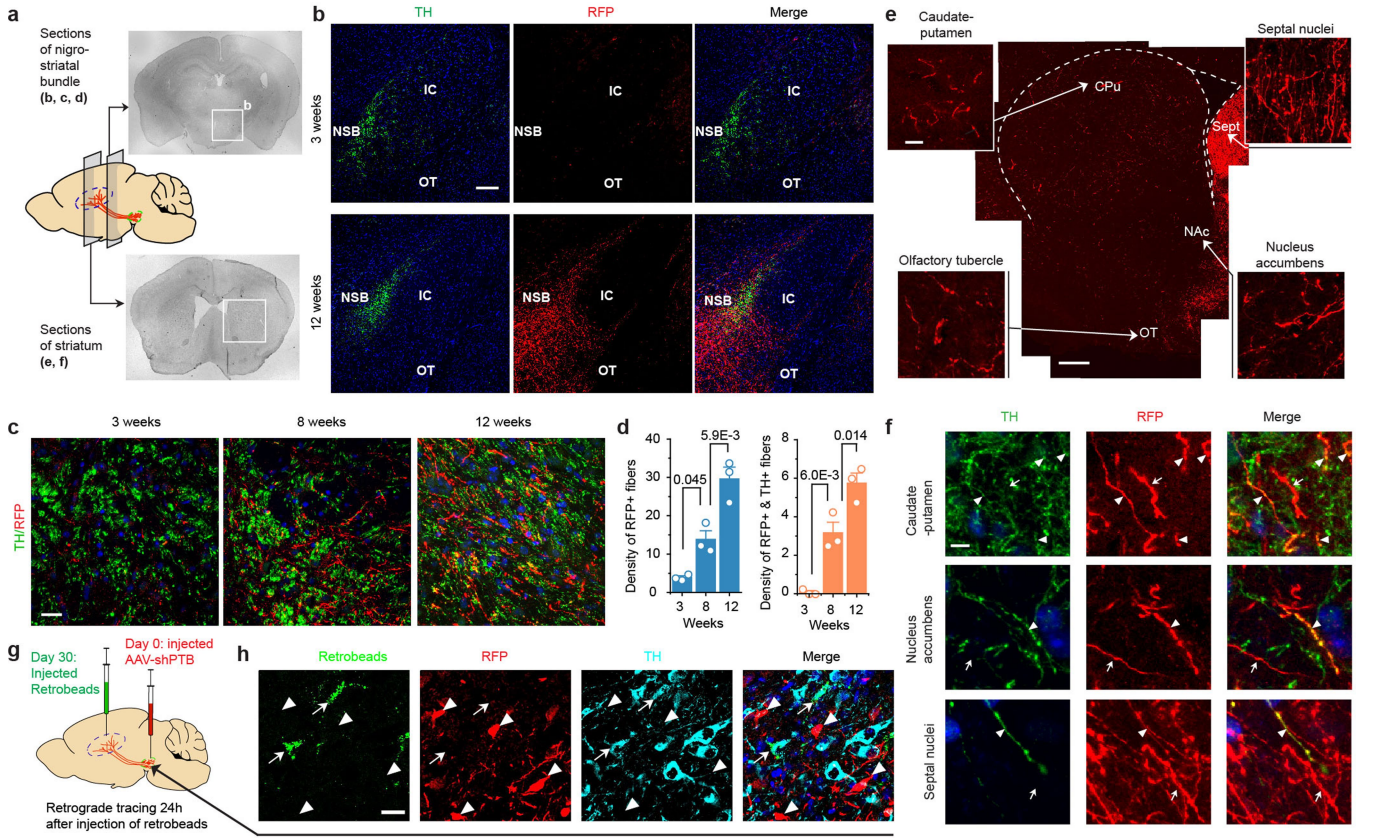
Extended Data Fig. 6 | Electrophysiological properties of gradually matured DA neurons. a, **b**, Schematic depiction of patch recording of converted neurons in midbrain (**a**). According to this scheme, the fluorescent dye Neurobiotin 488 (green) loaded in the electrode was used to mark cell bodies in substantia nigra for patch clamp recording on brain slices. **b**, After recording, the patched cells were confirmed to be RFP+TH+ to demonstrate the recording being performed on newly converted neurons (scale bar, 20 μm). Experiments were independently repeated 4 times with similar results. **c**–**e**, Detection of spontaneous action potential (**c**) and relatively wider action potential generated by newly converted neurons in comparison with endogenous GABAergic neurons (**d**). **e**, Notably, hyperpolarization-activated currents of HCN channels (I_h currents) were recorded at 12 weeks after, but not 6 weeks after, AAV-shPTB-induced neuronal conversion; these currents could be specifically blocked with CsCl. The numbers of cells that showed the recorded activity versus the total number of cells examined are indicated. Note that the bottom trace is also shown in Fig. 2h. **f**, **g**, Extracellular recording showing more converted neurons firing spontaneous action potentials at 12 weeks after transduction with AAV-shPTB than at 6 weeks after transduction. The numbers of cells that showed the recorded activity versus the total number of cells examined are indicated. **g**, The frequency of spontaneous spikes that increased upon further maturation was further quantified. Data were based on a total of 31 cells from 4 mice. Results show progressive maturation of newly converted DA neurons in the brain. Statistical significance was determined by two-sided Student's *t*-test. **h**, Cortical neurons generated in AAV-shPTB-transduced cortex, in contrast to a large population of RFP+TH+ cells in midbrain. As a control, AAV-shPTB was injected in cortex. After 12 weeks, RFP+ cells were co-stained with the cortical neuron marker CTIP2 (top) and CUX1 (bottom). Scale bars, 40 μm (main); 15 μm (magnified inset). Note that RFP+ CUX1+ cells are rare in comparison to RFP+CTIP2+ cells, indicative of different conversion efficiency in different layers of cortex. Experiments were independently repeated twice with similar results.



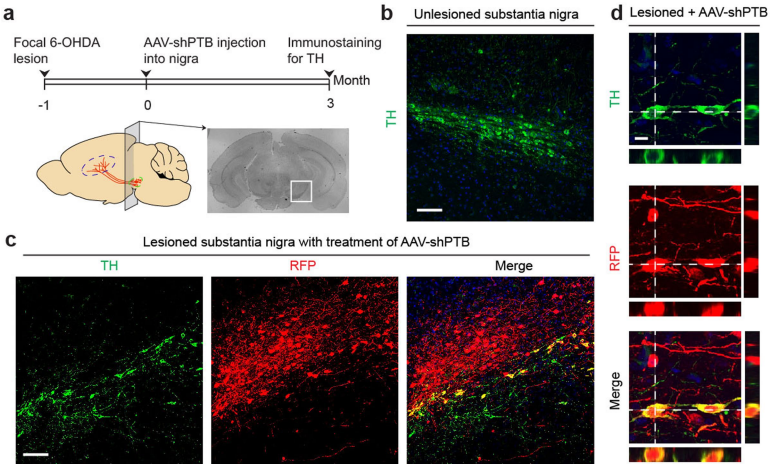
Extended Data Fig. 7 | Characterization of cortical astrocyte-derived neurons compared with midbrain astrocyte-derived neurons. a–**c**, A small fraction of cortical astrocyte-derived neurons express DA neuron markers. **a**, RT–qPCR showed the induction of DA neuron-specific genes *Slc6a3* and *Foxa2* in isolated cortical astrocytes treated with lentiviral shPTB. These DA-like neurons were further characterized by immunostaining for additional DA neuron markers DAT and VMAT2 (**b**; scale bar, 20 μm) and quantified among TUJ1+ cells based on 3 biological repeats with at least 100 cells counted in each (**c**). Two-sided Student's *t*-test; mean ± s.e.m. *P*-values are indicated. Results indicate that although cortex does not contain DA neurons and RFP+TH+ DA-like neurons were never detected in AAV-shPTB-transduced cortex in the brain, isolated cortical astrocytes were able to give rise to a fraction of DA-like neurons in vitro. This implies that astrocytes may become more plastic in culture than within specific brain environments. **d**, Additional immunochemical evidence for the expression of DA neuron-specific markers (LMX1A, PITX3 and DDC) in a subpopulation of TUJ1+ cells derived from cortical astrocytes. Scale bar, 20 μm. Experiments were independently repeated 3 times with similar results. **e**–**g**, TH staining of TUJ1+ neurons derived from midbrain astrocytes and comparison with neurons derived from cortical astrocytes. **e**, Lentiviral shPTB, but not control shRNA, converted midbrain astrocytes into TH+ DA neurons in culture. Scale bar, 25 μm. **f**, Conversion efficiencies of cortical and midbrain astrocytes, showing similar high percentage of TUJ1+ neurons (left), but a significantly higher percentage of DA neurons converted from midbrain astrocytes compared with cortical astrocytes (right). Data are based on 3 biological repeats with at least 200 cells counted in each. Statistical significance was determined by two-sided Student's *t*-test; mean ± s.e.m. *P*-values are indicated. **g**, Western blotting analysis of a pan-neuronal marker (TUJ1) and two specific markers for DA neurons (TH and VMAT2) in shPTB-reprogrammed astrocytes from cortex and midbrain, showing much higher levels of the DA neuron markers in neurons generated from midbrain astrocytes compared to cortical astrocytes. Experiments were independently repeated twice with similar results. Together, these data strongly suggest intrinsic cellular differences that are responsible for the generation of different neuron subtypes from astrocytes in different brain regions.



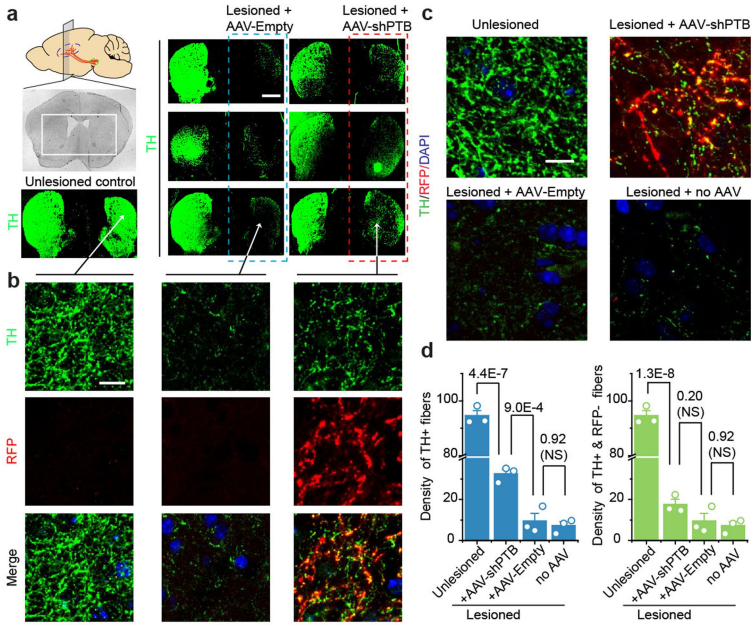
Extended Data Fig. 8 | Cell-autonomous mechanisms for the regional specificity in neuronal conversion. a, TH+ neurons generated from cortical astrocytes with normal and conditioned media from cultured midbrain astrocytes. Scale bar, 100 μm. **b**, Quantification of cells in **a**. Three biological repeats with at least 100 cells counted in each. Statistical significance was determined by ANOVA; mean ± s.e.m. **c**–**f**, RT–qPCR analysis of DA neuron-specific transcription factors in cortical and midbrain astrocytes before and after lentiviral shPTB-induced neuronal conversion. **c**, The indicated transcription factors were quantified by real-time PCR and normalized against β-actin mRNA. **d**, To ensure that the isolated astrocytes were free of contaminated neurons, RT–qPCR was also performed with the 3 indicated pan-neuron markers with isolated neurons as control. **e**, In response to PTB knockdown, astrocyte-specific genes *S100b* and *Gfap* were repressed, whereas pan-neuronal transcription factors *Myt1l* and *Ascl1* were activated in astrocytes derived from both cortex and midbrain. Dashed lines indicate levels before shPTB treatment, which was set to 1 for comparison with levels after shPTB treatment. **f**, Under the same conditions, the 4 DA-neuron-specific transcription factors were more robustly induced in response to PTB depletion in midbrain astrocytes compared to cortical astrocytes. Statistical significance was determined by ANOVA with post hoc Tukey test (**d**) or two-sided Student's *t*-test (**c**, **e**, **f**), based on 3 biological repeats; mean ± s.e.m. *P*-values are indicated. Results suggest higher basal levels and more robust induction of DA neuron-specific transcription factors in midbrain astrocytes compared to cortical astrocytes, providing evidence for the differences in cell-intrinsic gene expression programs in giving rise to distinct subtypes of neurons. **g**–**i**, Schematic of amperometric recording of monoamine release, showing the placement of a carbon fibre electrode on a midbrain astrocyte-derived neuron (**g**). Scale bar, 30 μm. **h**, Spike-like events were captured by holding the electrode at +750 mV after K^+ (25 mM) stimulation. **i**, A high-resolution view of dopamine release events in **h**. Results demonstrate a key functional property of midbrain astrocyte-derived DA neurons. Experiments were independently repeated twice with similar results.



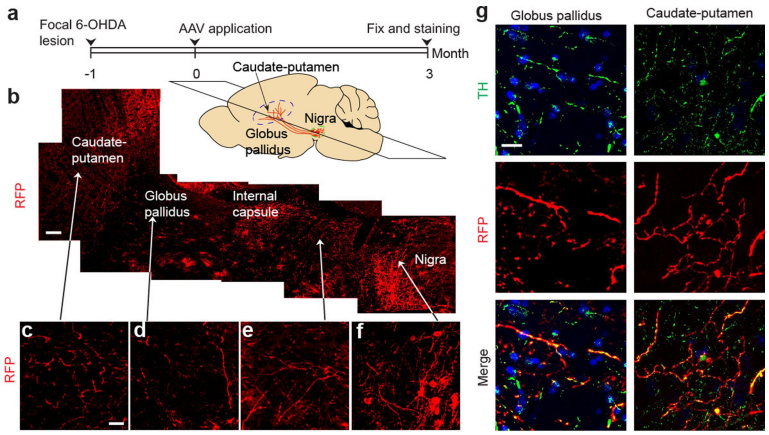
Extended Data Fig. 9 | Time-course analysis of fibre outgrowth from converted neurons. a, Schematic of coronal sections for analysing fibre density in the nigrostriatal pathway. **b**–**d**, Sphere-determined density of RFP+ fibres that were progressively increased along the nigrostriatal bundle (NSB). Shown are low-magnification views (**b**; scale bar, 150 µm) and enlarged views (**c**; scale bar, 35 µm). IC, internal capsule. **d**, Quantification of RFP+ (left) or RFP+TH+ fibres (right), based on 3 independent biological sections. Statistical significance was determined by ANOVA with post hoc Tukey test; mean ± s.e.m. *P*-values are indicated. Results show time-dependent increase in fibre density, a portion of which also exhibits colocalization of the DA neuron marker TH. **e**, Low-magnification view of striatum innervated by RFP+ projections. Scale bar, 300 µm. Smaller panels show magnified views of RFP+ projections in different regions. Scale bar, 15 µm. Note the bright RFP signals in septal nuclei. **f**, Three selected regions were further amplified to highlight a fraction of RFP+ fibres with (arrowheads) or without (arrows) co-staining with the DA neuron marker TH. Scale bar, 5 µm. Results emphasize that converted DA neurons targeted broader regions in striatum than endogenous DA neurons, which might cause side effects—a potential caveat of neuronal reprogramming experiments that requires investigation in future studies. **g**, **h**, Retrograde tracing of TH+ neurons from striatum to substantia nigra. Depicted is the AAV-shPTB injection site at day 0 and the retrobead injection site at day 30 (**g**). Retrograde tracing was monitored 24 h after injection of retrobeads. After treatment with AAV-shPTB for 30 days, TH+ cells, but not TH+RFP+ cells, in substantia nigra were labelled with retrograde beads (**h**). Arrowheads, RFP+ cells; arrows, cell bodies of endogenous TH+ DA neurons labelled with retrobeads. Scale bar, 20 µm. These data provide a critical control for AAV-shPTB-converted DA neurons that could be traced from striatum to substantia nigra, as described in the main text. All experiments shown in this figure were independently repeated 3 times with similar results.



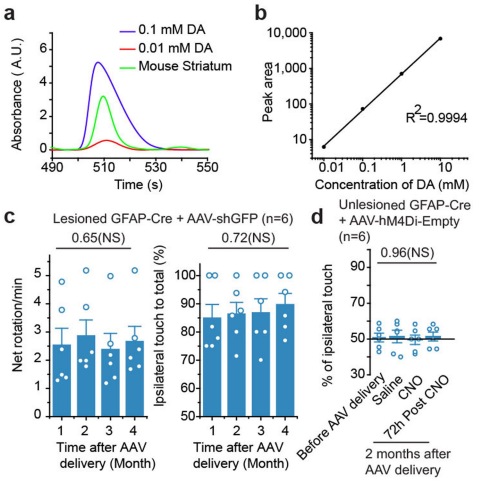
Extended Data Fig. 10 | shPTB-converted neurons replenish lost dopaminergic neurons in substantia nigra. a, Schematic of the experimental schedule for 6-OHDA-induced lesion followed by reprogramming with AAV-PTB and then TH staining. **b**, **c**, Low-magnification views of unlesioned substantia nigra stained for TH (**b**) and substantia nigra lesioned with 6-OHDA and transduced with AAV-shPTB (**c**). Scale bars, 80 μm. These data were used to provide the quantitative information shown in Fig. 4f, g. **d**, Enlarged view of RFP+ cells that co-expressed TH in substantia nigra. Two RFP+TH+ cell bodies are highlighted by orthogonal views of *z*-stacked images, attached on the right and bottom of the main image in each panel. Scale bar, 10 μm. Results show the generation of TH+ DA neurons in a highly region-specific manner in substantia nigra, as a large population of RFP+ cells were not labelled by TH staining in the same image. All experiments shown in this figure were independently repeated 3 times with similar results.



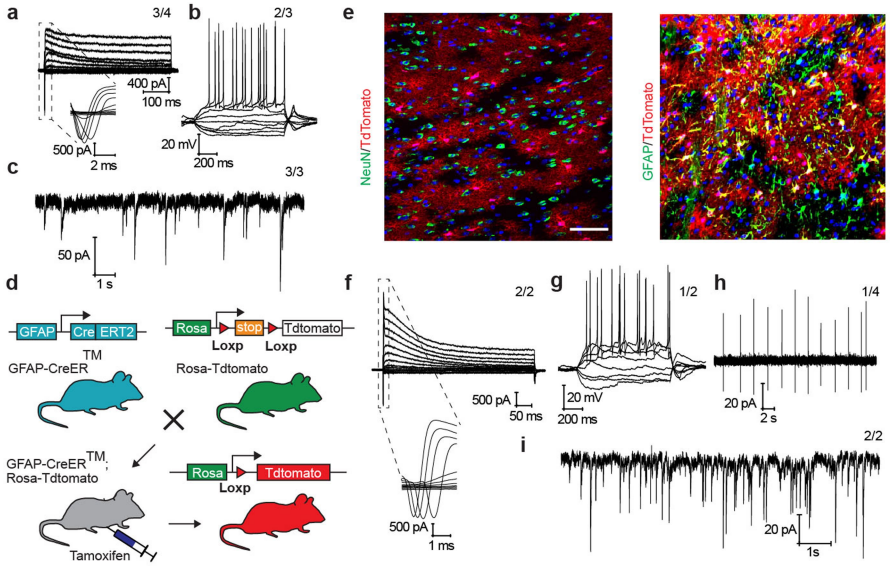
Extended Data Fig. 11 | Restoration of TH+ neurons in striatum of 6-OHDA-lesioned mice. a, **b**, Schematic of the coronal section of striatum and images of uninjured control and lesioned striatum treated on the right side of the brain with either AAV-empty or AAV-shPTB (**a**). Scale bar, 500 µm. **b**, Magnified images showed extensive colocalization of TH with RFP-labelled fibres. Scale bar, 10 µm. Results show a significant degree of restoration of TH+ fibres in striatum. Experiments were independently repeated 3 times with similar results. **c**, **d**, Quantitative analysis of TH+ fibres in striatum under different treatment conditions. TH staining of striatum under different treatment conditions, as indicated (**c**). Scale bar, 10 µm. **d**, Quantification of total TH+ or TH+RFP− fibre density in striatum under different treatment conditions based on 3 biological repeats. Statistical significance was determined by ANOVA with post hoc Tukey test; mean ± s.e.m. *P*-values are indicated. Results show that most TH+ fibres seem to derive from AAV-shPTB-converted dopaminergic neurons; however, the data do not rule out the possibility that the axons of some endogenous neurons also responded to the environment created by newly converted neurons.



Extended Data Fig. 12 | Reconstruction of the nigrostriatal pathway by converted dopaminergic neurons. a, Schematic of the experimental schedule for 6-OHDA-induced lesion and reconstruction of the nigrostriatal pathway. **b**–**f**, Images of RFP+ projections extending from substantia nigra to striatum. The schematic diagram shows the dorso-ventral level of the horizontal section. Scale bar, 100 µm. Magnified views show indicated brain regions (**c**–**f**). Scale bar, 25 µm. **g**, Amplified views of RFP-positive fibres that co-stained with TH in CPu and globus pallidus. Scale bar, 20 µm. These data were used to provide the quantitative information shown in Fig. 4h, i. Experiments were independently repeated twice with similar results.



Extended Data Fig. 13 | Measurement of striatal dopamine by HPLC and controls with AAV-shGFP and AAV-hM4Di. a, **b**, Dopamine levels in brain detected by HPLC with two different doses of spiked dopamine (**a**). **b**, Standard curve generated from the spiked dopamine. This set of experiments was performed only once. **c**, Controls for behavioural tests, showing that expressing an anti-GFP control shRNA alone did not rescue chemical-induced behavioural deficits based on apomorphine-induced rotation (left) and cylinder test (right). **d**, Controls for behavioural tests, showing that the expression of hM4Di in non-reprogrammed astrocytes did not trigger detectable behaviour change in non-lesioned mice in the presence of CNO. Statistical significance was determined by ANOVA (**c**, **d**); mean ± s.e.m. Six mice were analysed in each group. *P*-values are indicated.



Extended Data Fig. 14 | Electrophysiological analysis of PTB ASO-induced neurons in vitro and in brain. a–**c**, Converted neurons showed large currents from voltage-dependent sodium and potassium channels (**a**), repetitive action potentials (**b**) and spontaneous postsynaptic currents (**c**). The numbers of cells that showed the recorded activity versus the total number of cells examined are indicated on the top right in each panel. **d**, Schematic of transgenic mice used to trace astrocytes in vivo. **e**, Three weeks after tamoxifen treatment, none of the tdTomato-labelled cells in the midbrain of *Gfap-cre ER*:*Rosa-tdTomato* mice stained positive for NeuN (left), and all were GFAP+ (right). Scale bar, 50 µm. **f**–**i**, Converted neurons in brain slices showed large currents from voltage-dependent sodium and potassium channels (**f**), repetitive action potentials (**g**), spontaneous action potentials (**h**) and spontaneous postsynaptic currents (**i**). The numbers of cells that showed the recorded activity versus the total number of cells examined are indicated. The results show that functional neurons are induced by PTB ASO both in culture and in mouse brain. All experiments shown in this figure were independently repeated twice with similar results.

nature research

Corresponding author(s): Xiang-dong Fu

Last updated by author(s): Xiang-Dong Fu, Hao Qian

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Olympus FluoView Ver.4.2a, pClamp 10.0, Igor 4.04, MBA-1 DA/AD unit v4.07
Data analysis	Image-J 1.47v, Clampfit 10.4, OriginPro 2016, KMPlayer v4.0.7.1, MatLab v2009b, R v3.5.1, cutadapt v2.8, salmon v0.14.1, DEseq2 v1.22.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated or analyzed during this study are included in this published article (and its supplementary information files). The raw data from RNA-seq experiments have been deposited into NCBI GEO under the accession number GEO: GSE142250.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical analysis was employed to determine the sample size. The sample size was chosen base on previous publications using the same techniques.
Data exclusions	No data were excluded.
Replication	All key experiments in our manuscript have been repeated via intra- and inter-term collaboration and in different biological contexts. All statistical analysis were based on at least three replicates. More details of biological repeats are described in Supplementary Table 2.
Randomization	The mice used in all experiments were chosen randomly according to their unique identification numbers.
Blinding	Investigators were masked to group identity for measurements of striatal dopamine and dopamine release.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	We provide the all the essential information for antibodies in Supplementary Table 3.
Validation	The antibodies used were validated by the previous publications or routine experimental protocols, listed in Supplementary Table 3.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293T cells were from a common laboratory stock. Lenti-X 293T cells were purchased from Takara Bio (#632180). Mouse Embryonic Fibroblasts (MEF) were isolated from E14.5 C57BL/6 mouse embryos. Mouse neurons were isolated from E17~18 C57BL/6 mouse embryos. Mouse astrocytes were isolated from P4~P5 C57BL/6 mouse brain (see METHODS). Human dermal fibroblasts (HDF) were purchased from ATCC (PCS-201-012). Human neurons were trans-differentiated from human neuronal progenitor cells, which is a gift from Dr.Alysson Muotri's lab. Human astrocytes were purchased from Cell Applications (882AK-05f).
Authentication	The cell lines used were checked for morphology by microscopy and immunostaining with specific markers.
Mycoplasma contamination	We periodically checked potential contamination with mycoplasma, which causes retarded cell growth and low pH in the media. All cell lines tested negative for mycoplasma contamination by Hoechst staining of the cells according to Young L.et al., Nature Protocols,2010.
Commonly misidentified lines (See ICLAC register)	No. The cell lines used are not listed in the database.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	All animal experiments were conducted in accordance with the guide of The University of California San Diego Institutional~40

Laboratory animals	Animal Care and Use Committee (Protocol# S99116). Transgenic mice including B6.Cg-Tg(Gfap-cre)77.6Mvs/2J, B6.SJLSlc6a3tm1.1(cre)Bkmn/J, B6.Cg-Tg(GFAP-cre/ERT2)505Fmv/J, and B6.Cg-Gt(ROSA)26Sortm14(CAG-tdTomato)Hze/J were purchased from The Jackson Laboratory. Both male and female mice were used in this study. All mice at age of postnatal day 30~40 were used. 1year old B6.Cg-Tg(Gfap-cre)77.6Mvs/2J mice were also used to perform surgery to induce lesion.
Wild animals	This study did not involve wild animals.
Field-collected samples	This study did not involve samples collected from the field.
Ethics oversight	All procedures were conducted in accordance with the guide of The University of California San Diego Institutional Animal Care and Use Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Aerodynamic analysis of SARS-CoV-2 in two Wuhan hospitals

https://doi.org/10.1038/s41586-020-2271-3

Received: 14 March 2020

Accepted: 20 April 2020

Published online: 27 April 2020

Check for updates

Yuan Liu[1,6], Zhi Ning[2,6]✉, Yu Chen[1,6]✉, Ming Guo[1,6], Yingle Liu[1], Nirmal Kumar Gali[2], Li Sun[2], Yusen Duan[3], Jing Cai[4], Dane Westerdahl[2], Xinjin Liu[1], Ke Xu[1], Kin-fai Ho[5]✉, Haidong Kan[4]✉, Qingyan Fu[3]✉ & Ke Lan[1]✉

The ongoing outbreak of coronavirus disease 2019 (COVID-19) has spread rapidly on a global scale. Although it is clear that severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is transmitted through human respiratory droplets and direct contact, the potential for aerosol transmission is poorly understood[1–3]. Here we investigated the aerodynamic nature of SARS-CoV-2 by measuring viral RNA in aerosols in different areas of two Wuhan hospitals during the outbreak of COVID-19 in February and March 2020. The concentration of SARS-CoV-2 RNA in aerosols that was detected in isolation wards and ventilated patient rooms was very low, but it was higher in the toilet areas used by the patients. Levels of airborne SARS-CoV-2 RNA in the most public areas was undetectable, except in two areas that were prone to crowding; this increase was possibly due to individuals infected with SARS-CoV-2 in the crowd. We found that some medical staff areas initially had high concentrations of viral RNA with aerosol size distributions that showed peaks in the submicrometre and/or supermicrometre regions; however, these levels were reduced to undetectable levels after implementation of rigorous sanitization procedures. Although we have not established the infectivity of the virus detected in these hospital areas, we propose that SARS-CoV-2 may have the potential to be transmitted through aerosols. Our results indicate that room ventilation, open space, sanitization of protective apparel, and proper use and disinfection of toilet areas can effectively limit the concentration of SARS-CoV-2 RNA in aerosols. Future work should explore the infectivity of aerosolized virus.

The ongoing outbreak of COVID-19, which has been reported in 206 countries and areas, has resulted in 857,641 confirmed cases and 42,006 deaths globally as of 2 April 2020. Owing to the increasing threat caused by COVID-19 to global health, the World Health Organization (WHO) has declared the COVID-19 outbreak a pandemic and global public health emergency. The causative pathogen of the COVID-19 outbreak has been identified as a highly infectious novel coronavirus that is referred to as SARS-CoV-2[4–6]. Reported transmission pathways of SARS-CoV-2 in humans include the inhalation of virus-laden liquid droplets, close contact with infected individuals and contact with surfaces that are contaminated with SARS-CoV-2[1]. Moreover, aerosol transmission has been suggested to be an additional, yet important pathway, on the basis of clinical observations in confined spaces[2,3]. There are many respiratory diseases that are spread through airborne routes, such as tuberculosis, measles and chickenpox[7,8]. A retrospective cohort study conducted after the SARS epidemic—which was caused by SARS-CoV—in Hong Kong in 2003 suggested that airborne spread may have had an important role in the transmission of SARS[9]. At present, little is known about the aerodynamic characteristics and transmission pathways of SARS-CoV-2 in aerosols; in part because of the difficulties in sampling virus-containing aerosols in real-world settings and challenges in their quantification at low concentrations.

We analysed the occurrence of airborne SARS-CoV-2 and its aerosol deposition at 30 sites in two designated hospitals and public areas in Wuhan, China, and then quantified the copy counts of SARS-CoV-2 in aerosol samples using a robust droplet-digital-PCR-based detection method (ddPCR)[10]. The two hospitals are exclusively used for the treatments of patients with COVID-19 during the outbreak; however, each hospital has unique characteristics that serve different purposes. Renmin Hospital of Wuhan University (hereafter, Renmin Hospital) is representative of grade-A tertiary hospitals that have been designated for the treatment of patients with severe symptoms of COVID-19. By contrast, Wuchang Fangcang Field Hospital (hereafter, Fangcang Hospital) is representative of the makeshift field hospitals that were converted from indoor sports facilities or exhibition centres to quarantine and treat patients with mild symptoms. The sampling locations

[1]State Key Laboratory of Virology, Modern Virology Research Center, College of Life Sciences, Wuhan University, Wuhan, P. R. China. [2]Division of Environment and Sustainability, The Hong Kong University of Science and Technology, Hong Kong, P. R. China. [3]Shanghai Environmental Monitoring Center, Shanghai, P. R. China. [4]School of Public Health, Key Laboratory of Public Health Safety of the Ministry of Education and Key Laboratory of Health Technology Assessment of the Ministry of Health, Fudan University, Shanghai, P. R. China. [5]JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong, P. R. China. [6]These authors contributed equally: Yuan Liu, Zhi Ning, Yu Chen, Ming Guo. ✉e-mail: zhining@ust.hk; chenyu@whu.edu.cn; kfho@cuhk.edu.hk; kanh@fudan.edu.cn; qingyanf@sheemc.cn; klan@whu.edu.cn

Table 1 | RNA concentration of airborne SARS-CoV-2 at different locations in Wuhan

Category	Sites	Sample type	Concentration (copies m^{-3})
Patient areas			
Fangcang Hospital	Zone A workstation[a]	TSP[b]	1
		TSP[c]	9
	Zone B workstation	TSP	1
	Zone C workstation[a]	TSP[b]	5
		TSP[c]	0
	Patient mobile toilet room	TSP	19
Renmin Hospital	Intensive care unit	TSP	0
	Intensive care unit	Deposition	31[d]
	Intensive care unit	Deposition	113[d]
	Coronary care unit	TSP	0
	Ward zone 16	TSP	0
Medical staff areas			
Fangcang Hospital	PPAR of zone A[a]	TSP[b]	16
		TSP[c]	0
	PPAR of zone B	Size-segregated	42
	PPAR of zone C[a]	Size-segregated[b]	20
		TSP[c]	0
	Male staff change room	TSP	20
	Female staff change room	TSP	11
	Medical staff's office	Size-segregated	20
	Meeting room	TSP	18
	Warehouse[a]	TSP[b]	21
		TSP[c]	0
Renmin Hospital	Passageway for medical staff	TSP	6
	Dining room for medical staff	TSP	6
Public areas			
	Fangcang Hospital pharmacy	TSP	3
	Renmin Hospital doctor office	TSP	0
	Renmin Hospital outpatient hall	TSP	0
	Renmin Hospital outdoor	TSP	7
	University office doorside	TSP	0
	University hospital outpatient hall	TSP	0
	Community checkpoint	TSP	0
	Residential building	TSP	0
	Supermarket	TSP	0
	Department store 1	TSP	11
	Department store 2	TSP	3
	Blank control[a]	Field blank[b]	0
		Field blank[c]	0

TSP, total suspended particles. The samples were distinct by design owing to the unique conditions inside the hospitals during COVID-19 outbreak. We collected 35 samples (not including two blanks) at different sites, therefore $n = 35$. The replicability is limited by very restricted experimental conditions to conduct sampling in the highly infectious zones.

[a]Two rounds of sampling were conducted for the sites. A blank control was included for each of round of sampling. Detailed information is shown in Supplementary Table 1.

[b]The samples were taken during the first round of sampling from 17 February to 24 February 2020.

[c]The samples were taken during the second round of sampling on 2 March 2020.

[d]The reported values are virus aerosol deposition rates in copies m^{-2} h^{-1}.

were classified into three categories according to their accessibility by different groups: (1) patient areas, where the patients with COVID-19 have a physical presence—these include the intensive care units, coronary care units and ward rooms inside Renmin Hospital, a toilet and staff workstations inside Fangcang Hospital; (2) medical staff areas, the workplaces in the two hospitals that are exclusively accessed by medical staff who had direct contact with the patients; and (3) public areas, venues that are open to the general public (Supplementary Table 1). Three types of aerosol samples were collected: (1) aerosol samples of total suspended particles with no upper size limit to quantify RNA concentrations of SARS-CoV-2 in aerosols; (2) aerodynamic size-segregated aerosol samples to determine the size distribution of airborne SARS-CoV-2 droplets; and (3) aerosol deposition samples to determine the deposition rate of airborne SARS-CoV-2.

The existence of SARS-CoV-2 in aerosol samples was determined through the quantification of its genetic material (RNA).

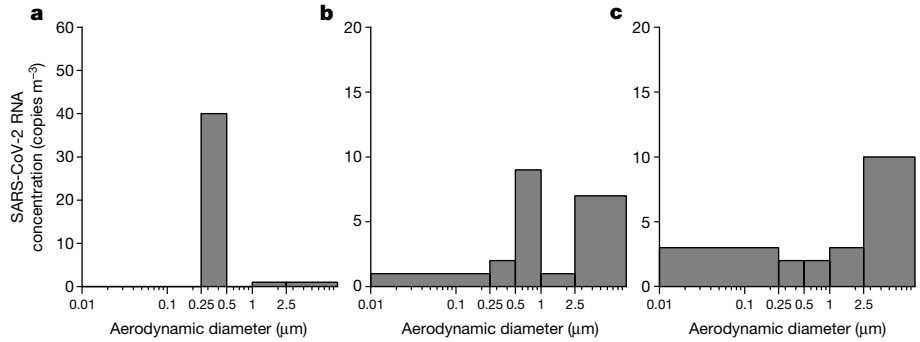


Fig. 1 | Concentration of airborne SARS-CoV-2 RNA in different aerosol size bins. a, Concentration of SARS-CoV-2 in a protective-apparel removal room in zone B of Fangcang Hospital. **b**, Concentration of SARS-CoV-2 in a protective-apparel removal room in zone C of Fangcang Hospital. **c**, Concentration of SARS-CoV-2 in the medical staff's office of Fangcang Hospital. The *x* axis represents the aerodynamic diameter on a logarithmic scale to cover the multiple magnitudes of measured aerosol diameters.

The concentrations of airborne SARS-CoV-2 at the different sites are shown in Table 1. In general, very low or undetectable concentrations of airborne SARS-CoV-2 were found in most of the patient areas of Renmin Hospital, suggesting that the negatively pressurized isolation and high air exchange rate inside the intensive care units, coronary care units and ward room of Renmin Hospital are very effective in limiting the airborne transmission of SARS-CoV-2. The highest concentration in patient areas was observed inside a patient mobile toilet room at Fangcang Hospital (19 copies m^{-3}), which is a temporary single toilet room of approximate 1 m^2 in area without ventilation. Airborne SARS-CoV-2 may come from either the patient's breath or the aerosolization of the virus-laden aerosol from the faeces or urine of a patient during use[11,12]. Although the infectivity of the virus is not known in this study, the results also relate to the findings of another study[13], which found positive test results of wipe samples from room surfaces of toilets used by patients infected with SARS-CoV-2. In medical staff areas, the two sampling sites in Renmin Hospital had low concentrations of 6 copies m^{-3}, whereas the sites in Fangcang Hospital generally had higher concentrations. In particular, the protective-apparel removal rooms (PPARs) in three different zones inside Fangcang Hospital are among the upper range of the concentrations of airborne SARS-CoV-2, ranging from 16 to 42 copies m^{-3} in the first round of sampling. In public areas outside the hospitals, we found that most of the sites had undetectable or very low concentrations of SARS-CoV-2 aerosols (below 3 copies m^{-3}), except for one crowd-gathering site about 1 m from the entrance of a department store that customers frequently passed through and a site next to Renmin Hospital, through which the public including outpatients walked. Although both sites were outside buildings, it is possible that individuals infected with SARS-CoV-2 in the crowd may have been the source of virus-laden aerosols during the sampling period. The results suggest that, overall, the risks of infection are low in well-ventilated or open public venues, but do reinforce the importance of avoiding crowded gatherings and implementing the early identification and diagnosis of individuals infected with SARS-CoV-2 for quarantine or treatment.

Inside a room of the intensive care unit of Renmin Hospital, the two aerosol deposition samples tested positive with an estimated deposition rate of 31 and 113 copies m^{-2} h^{-1}, although the concentration of the total suspended particles in the aerosol sample inside this room of the intensive care unit was below the detection limit (Table 1). The sample with the higher deposition rate was placed in the hindrance-free corner of the room, approximately 3 m from the bed of a patient. The other sample, for which a lower number of virus copies was recorded, was placed in another corner, approximately 2 m from the bed of the patient and below medical equipment, which may have blocked the path of virus aerosols during sedimentation. Our findings, although based on a small sample size, indicate that virus-laden aerosol deposition may have a role in surface contamination and subsequent contact by susceptible people, which results in the infection of individuals with SARS-CoV-2.

In general, medical staff areas had higher concentrations of SARS-CoV-2 aerosols compared with patient areas in both hospitals during the first round of sampling (17–24 February 2020) at the peak of the COVID-19 outbreak (Table 1). For sampling sites at Renmin Hospital, the air circulation in medical staff areas is isolated by design from the air circulation in the patient rooms. By contrast, in Fangcang Hospital, the non-ventilated temporary PPAR was isolated from the patient hall, in which the aerosol concentration of SARS-CoV-2 was generally low. The second round of sampling of total suspended particles in medical staff areas of Fangcang Hospital was conducted after the number of patients reduced from more than 200 to less than 100 per zone and the implementation of more rigorous and thorough sanitization measures, including more frequent spraying of chlorinated disinfectant on the floor of patient areas, additional disinfection using 3% hydrogen peroxide in the PPAR at least once a week, thoroughly spraying alcohol disinfectant on the protective apparel before taking it off and an increased operation time of indoor air purifiers. The samples from this second round showed all undetectable results (Table 1), confirming the importance of sanitization in reducing the amount of airborne SARS-CoV-2 in high-risk areas.

SARS-CoV-2 aerosols were mainly found to include two size ranges, one in the submicrometre region (d_p between 0.25 and 1.0 μm) and the other in supermicrometre region (d_p > 2.5 μm). Aerosols in the submicrometre region were predominantly found in PPARs in zones B and C of Fangcang Hospital (Fig. 1a, b) with peak concentrations of 40 and 9 copies m^{-3} in the 0.25–0.5 μm and 0.5–1.0 μm range, respectively. By contrast, aerosols in the supermicrometre region were mainly observed in the PPAR of zone C of Fangcang Hospital (Fig. 1b) with concentrations of 7 copies m^{-3}. The medical staff's office (Fig. 1c) had more virus-laden aerosols in the supermicrometre size range, but the size distribution is flatter compared with the range in other areas. Reports on the resuspension of microorganisms from the floor, clothing and furniture have previously been noted to contribute to the generation of microbial aerosols in the built environment[14]. Therefore, we hypothesize that the source of the submicrometre peak is the resuspension of virus-laden aerosols from the surface of the protective apparel worn by medical staff while they are removing the equipment. The submicrometre virus-laden aerosols may originally come from the direct deposition of respiratory droplets or airborne SARS-CoV-2 from a patient onto the protective apparel as evidenced by the deposition samples (Table 1). The higher mobility owing to their smaller aerodynamic diameter facilitates the resuspension from the surface of protective

apparel after gaining the initial velocity while the equipment is being removed. On the other hand, floor-deposited SARS-CoV-2 is possibly the source of supermicrometre virus-laden aerosols and was carried across different areas by medical staff. Furthermore, a recent study has experimentally demonstrated that SARS-CoV-2 could maintain its biological stability in aerosols and on different surfaces for hours to days[15]. The submicrometre SARS-CoV-2 aerosols found in this study had a relatively longer residence time, indicating that the virus was probably still infectious during transmission.

This study has its inherent limitations because of the small sample size and the description of sample viral RNA instead of virus infectivity, which was imposed by restricted access to the patient and medical staff areas at the epicentre of the COVID-19 outbreak. Nonetheless, the findings of this study provide a real-world investigation of the aerodynamic characteristics of airborne SARS-CoV-2 in Wuhan, where a strict quarantine and travel restrictions were implemented during the peak of the COVID-19 outbreak. The findings suggest that toilet use by patients with COVID-19 and crowd gatherings that included individuals infected by SARS-CoV-2 are non-negligible sources of airborne SARS-CoV-2, although the infectivity of the virus is not known. We also describe a transmission pathway for SARS-CoV-2 aerosols that is mediated by the surface deposition of the virus on and resuspension from protective apparel of medical staff and the floor surface. The results of this study have important implications for the prevention of infection of the public and protection of medical staff. We call for particular attention to (1) the ventilation and sterilization of toilets as a potential source for the spreading of the virus; (2) personal protection measures for the general public, such as the wearing of masks and avoidance of busy crowds to reduce the risk of exposure to airborne virus; (3) the effective sanitization of high-risk areas in the hospital to limit the transmission of airborne SARS-CoV-2 and to protect the medical staff; (4) the effectiveness of a naturally ventilated large stadium to limit the aerosol transmission of SARS-CoV-2 when converted to a field hospital for the quarantine and treatment of patients with SARS-CoV-2; and (5) surface sanitization of the apparel before the equipment is taken off to help to reduce the potential risk of infection for medical staff.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2271-3.

1. National Center for Immunization and Respiratory Diseases (NCIRD). *How COVID-19 spreads* https://www.cdc.gov/coronavirus/2019-ncov/about/transmission.html (2020).
2. Offord, C. How COVID-19 is spread. *The Scientist* https://www.the-scientist.com/news-opinion/how-covid-19-is-spread-67143 (21 February 2020).
3. WHO *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)* https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf (2020).
4. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species *Severe acute respiratory syndrome-related coronavirus*: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5**, 536–544 (2020).
5. Chen, L. et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. *Emerg. Microbes Infect.* **9**, 313–319 (2020).
6. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
7. Leclair, J. M., Zaia, J. A., Levin, M. J., Congdon, R. G. & Goldmann, D. A. Airborne transmission of chickenpox in a hospital. *N. Engl. J. Med.* **302**, 450–453 (1980).
8. Escombe, A. R. et al. The detection of airborne transmission of tuberculosis from HIV-infected patients, using an in vivo air sampling model. *Clin. Infect. Dis.* **44**, 1349–1357 (2007).
9. Yu, I. T., Qiu, H., Tse, L. A. & Wong, T. W. Severe acute respiratory syndrome beyond Amoy Gardens: completing the incomplete legacy. *Clin. Infect. Dis.* **58**, 683–686 (2014).
10. Suo, T. et al. ddPCR: a more accurate tool for SARS-CoV-2 detection in low viral load specimens. *Emerg. Microbes Infect.* https://doi.org/10.1080/22221751.2020.1772678 (2020).
11. Peng, L. et al. SARS-CoV-2 can be detected in urine, blood, anal swabs, and oropharyngeal swabs specimens. *J. Med. Virol.* https://doi.org/10.1002/jmv.25936 (2020).
12. Wang, W. et al. Detection of SARS-CoV-2 in different types of clinical specimens. *J. Am. Med. Assoc.* **323**, 1843–1844 (2020).
13. Ong, S. W. X. et al. Air, surface environmental, and personal protective equipment contamination by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) from a symptomatic patient. *J. Am. Med. Assoc.* **323**, 1610–1612 (2020).
14. Prussin, A. J. II & Marr, L. C. Sources of airborne microorganisms in the built environment. *Microbiome* **3**, 78 (2015).
15. van Doremalen, N. et al. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *N. Engl. J. Med.* **382**, 1564–1567 (2020).

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Sample collection

The sampling was conducted between 17 February and 2 March 2020 in the locations in two rounds as shown in Table 1. All aerosol samples were collected on presterilized gelatin filters (Sartorius). A total of 30 aerosol samples of total suspended particles were collected on 25-mm-diameter filters loaded into styrene filter cassettes (SKC) by sampling air at a fixed flow rate of 5.0 l min^{-1} using a portable pump (APEX2, Casella). A total of three size-segregated aerosol samples was collected using a miniature cascade impactor (Sioutas Impactor, SKC) that separated aerosols into five ranges (>2.5 μm, 1.0–2.5 μm, 0.50–1.0 μm and 0.25–0.50 μm on 25-mm filter substrates, and 0–0.25 μm on 37-mm filters) at a flow rate of 9.0 l min^{-1}. A total of two aerosol deposition samples was collected using 80-mm-diameter filters packed into a holder with an effective deposition area of 43.0 cm^2 and the filters were placed intact on the floor in two corners of the intensive care unit room of Renmin Hospital for 7 days. Sampling durations and operation periods are described in Supplementary Table 1. All sampling instruments were located in the centre of the respective sampling area, where the sampling inlet was at a height of 1.5 m from the floor. Considering the limited experimental conditions and the small sample size, the integrity and robustness of the experiment protocol was examined extensively in the laboratory before field sampling and these results are described in Supplementary Table 2.

Analytical method and data analysis

After the collection of aerosol samples, all samples were processed immediately in the BSL-2 laboratory of Wuhan University. The 25-, 37-mm and 80-mm filter samples were dissolved in deionized water, after which TRIzol LS reagent (Invitrogen) was added to inactivate SARS-CoV-2 viruses and extract RNA according to the manufacturer's instructions. First-strand cDNA was synthesized using the PrimeScript RT kit (TakaRa). Optimized ddPCR was used to detect the presence of SARS-CoV-2 viruses according to a previous study[10]. Analysis of the ddPCR data was performed using QuantaSoft software (Bio-Rad). The concentration reported by the procedure equals the number of copies of template per microlitre of the final 1× ddPCR reaction, which was normalized to copies m^{-3} in all of the results; therefore, the virus or viral RNA concentration in aerosol is expressed in copies m^{-3} throughout. A detailed protocol is provided in the Supplementary Information.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data generated and analysed during this study are included in the Article and its Supplementary Information. Source Data for Fig. 1 are provided with the paper.

Acknowledgements This study was supported by the Special Fund for COVID-19 Research of Wuhan University. We thank Taikang Insurance Group, Beijing Taikang Yicai Foundation, Renmin Hospital and Wuchang Fangcang Hospital for their support of this work; H. Xu from Xi'an Jiaotong University, Qingdao Laoying Environmental Technology, Beijing Top Science, Shanghai Leon Scientific Instrument, Shanghai Eureka Environmental Protection Hi-tech and Sapiens Environmental Technology for their support in providing the sampling devices and technical support of this study; and C. Wang, Q. Zhang, G. Liang and Z. Song for their assistance in filter sample preparation and logistics support.

Author contributions K.L., Y.C., Z.N., Q.F., H.K. and K.-f.H. conceptualized the study design; Yuan Liu, Y.C., M.G., Yingle Liu and K.L. collected samples; Yuan Liu, M.G. and X.L. carried out the laboratory tests; Yuan Liu, Z.N., Y.C., N.K.G., M.G., X.L. and K.L. analysed the data; Y.C., Z.N., Yuan Liu, Q.F., H.K., J.C., K.-f.H. and K.L. interpreted the results; Yuan Liu and Z.N. wrote the initial drafts of the manuscript; Yuan Liu, Z.N., Y.C. and K.L. revised the manuscript; M.G., Yingle Liu, N.K.G., L.S., Y.D., J.C., D.W., K.X., H.K. and Q.F. commented on the manuscript. All authors read and approved the final manuscript.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2271-3.
Correspondence and requests for materials should be addressed to Z.N., Y.C., K.-f.H., H.K., Q.F. or K.L.
Peer review information *Nature* thanks Linsey Marr and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.
Reprints and permissions information is available at http://www.nature.com/reprints.

nature research

Corresponding author(s): Ke Lan

Last updated by author(s): Apr 2, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	NONE
Data analysis	QuantaSoft analysis software v.1.7.4.0917 (Bio-Rad) was used for analysis of the ddPCR data.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated and analysed during this study are included in the Article and its Supplementary Information files, which include an additional Excel file containing Source Data for Fig. 1 and all raw data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study investigated the aerodynamic nature and aerosol transmission of SARS-CoV-2 aerosol in Wuhan under strict quarantine and travel restriction during the peak of COVID-19 outbreak.
Research sample	Total of 35 distinct samples of three types, total suspended particle, size segregated and aerosol deposition were collected.
Sampling strategy	We sampled three types of virus aerosol samples at 30 sites covering patient and medical staff areas inside hospitals and in public areas in Wuhan. The sampling was designed to identify the hotspots of airborne SARS-CoV-2 and investigate their sources, and seek for evidences of their aerosol transmission across different isolation zones and air-surface transfer mechanisms.
Data collection	We collected two batches of samples from Feb 17 to Feb 24, 2020 and on Mar 2, 2020 respectively.
Timing and spatial scale	Sampling durations range from 5 to 20 hours for total suspended particle and size segregated samples and 7 days for aerosol deposition samples. Sampling air volumes range from 1.5 m3 to 8.9 m3.
Data exclusions	No data were excluded in this work.
Reproducibility	The samples were distinct by design in this study due to the unique conditions inside the hospitals during COVID-19 outbreak.
Randomization	This study categorized sampling locations by functions and user groups. Each sample has served different purposes so no randomization was attempted.
Blinding	The sampling process itself has no impact on the study subject and data integrity by the nature of study, so no blinding was attempted.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	The sampling sites include indoor of hospital function areas and outdoor in public areas in Wuhan under strict quarantine and travel restriction.
Location	We sampled SARS-CoV-2 aerosol samples at 30 sites in two designated hospitals and public areas in Wuhan.
Access and import/export	This field study didn't involve any study objects that require permission so no approval is needed.
Disturbance	The sampling process has no disturbance of the subjects.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Article

Rapid reconstruction of SARS-CoV-2 using a synthetic genomics platform

https://doi.org/10.1038/s41586-020-2294-9

Received: 20 February 2020

Accepted: 24 April 2020

Published online: 4 May 2020

Check for updates

Tran Thi Nhu Thao[1,2,3,10], Fabien Labroussaa[2,4,10], Nadine Ebert[1,2,10], Philip V'kovski[1,2], Hanspeter Stalder[1,2], Jasmine Portmann[1,2], Jenna Kelly[1,2], Silvio Steiner[1,2,3], Melle Holwerda[1,2,3,5], Annika Kratzel[1,2,3], Mitra Gultom[1,2,3,5], Kimberly Schmied[1,2], Laura Laloli[1,2,3,5], Linda Hüsser[1,2], Manon Wider[5], Stephanie Pfaender[1,2,6], Dagny Hirt[1,2], Valentina Cippà[2,4], Silvia Crespo-Pomar[2,4], Simon Schröder[7], Doreen Muth[7,8], Daniela Niemeyer[7,8], Victor M. Corman[7,8], Marcel A. Müller[7,8,9], Christian Drosten[7,8], Ronald Dijkman[1,2,5], Joerg Jores[2,4,11] ✉ & Volker Thiel[1,2,11] ✉

Reverse genetics has been an indispensable tool to gain insights into viral pathogenesis and vaccine development. The genomes of large RNA viruses, such as those from coronaviruses, are cumbersome to clone and manipulate in *Escherichia coli* owing to the size and occasional instability of the genome[1–3]. Therefore, an alternative rapid and robust reverse-genetics platform for RNA viruses would benefit the research community. Here we show the full functionality of a yeast-based synthetic genomics platform to genetically reconstruct diverse RNA viruses, including members of the *Coronaviridae*, *Flaviviridae* and *Pneumoviridae* families. Viral subgenomic fragments were generated using viral isolates, cloned viral DNA, clinical samples or synthetic DNA, and these fragments were then reassembled in one step in *Saccharomyces cerevisiae* using transformation-associated recombination cloning to maintain the genome as a yeast artificial chromosome. T7 RNA polymerase was then used to generate infectious RNA to rescue viable virus. Using this platform, we were able to engineer and generate chemically synthesized clones of the virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[4], which has caused the recent pandemic of coronavirus disease (COVID-19), in only a week after receipt of the synthetic DNA fragments. The technical advance that we describe here facilitates rapid responses to emerging viruses as it enables the real-time generation and functional characterization of evolving RNA virus variants during an outbreak.

Within the past decade, we have seen outbreaks of numerous viruses, including Middle East respiratory syndrome coronavirus (MERS-CoV)[5], ZIKA virus[6], Ebola virus[7] and, at the end of 2019, SARS-CoV-2—which was first detected in Wuhan, Hubei province, China[4], but rapidly developed into a pandemic. During the early phase of the SARS-CoV-2 outbreak, virus isolates were not available to health authorities and the scientific community, even though these isolates are urgently needed to generate diagnostic tools, to develop and assess antivirals and vaccines, and to establish appropriate in vivo models. The generation of the SARS-CoV-2 from chemically synthesized DNA could bypass the limited availability of virus isolates and would furthermore enable genetic modifications and functional characterization. However, although *E. coli* proved very useful for the cloning of many viral genomes, it has a number of disadvantages when used for the assembly and stable maintenance of full-length molecular clones of emerging RNA viruses, including coronaviruses.

Synthetic genomics is a field fuelled by the efforts to create a bacterial cell that is controlled by a synthetic genome[8]. Genome-wide reassembly of the approximately 1.1-megabase (Mb) genome of *Mycoplasma* was first attempted using *E. coli* as an intermediate host[8]; however, the maintenance of 100-kilobase (kb) DNA fragments appeared to be very difficult in this host. Therefore, the yeast *S. cerevisiae* was chosen to clone, assemble and mutagenize entire *Mycoplasma* genomes[9,10]. The rationale for using a yeast cloning system is the ability of yeast to recombine overlapping DNA fragments in vivo, which led to the development of a technique called transformation-associated recombination (TAR) cloning[11].

More recently[12,13], TAR cloning was successfully used for the assembly, genetic engineering and rescue of large DNA viruses such as cytomegalovirus and herpes simplex virus 1. For coronaviruses that belong to a family of positive-stranded RNA viruses termed *Coronaviridae*, the generation of full-length molecular clones has long been hampered by

[1]Institute of Virology and Immunology (IVI), Bern, Switzerland. [2]Department of Infectious Diseases and Pathobiology, Vetsuisse Faculty, University of Bern, Bern, Switzerland. [3]Graduate School for Biomedical Science, University of Bern, Bern, Switzerland. [4]Institute of Veterinary Bacteriology, Vetsuisse Faculty, University of Bern, Bern, Switzerland. [5]Insitute for Infectious Diseases, University of Bern, Bern, Switzerland. [6]Department for Molecular and Medical Virology, Ruhr-Universität Bochum, Bochum, Germany. [7]Institute of Virology, Charité-Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany. [8]German Centre for Infection Research, associated partner Charité, Berlin, Germany. [9]Martsinovsky Institute of Medical Parasitology, Tropical and Vector Borne Diseases, Sechenov University, Moscow, Russia. [10]These authors contributed equally: Tran Thi Nhu Thao, Fabien Labroussaa, Nadine Ebert. [11]These authors jointly supervised this work: Joerg Jores, Volker Thiel. ✉e-mail: joerg.jores@vetsuisse.unibe.ch; volker.thiel@vetsuisse.unibe.ch

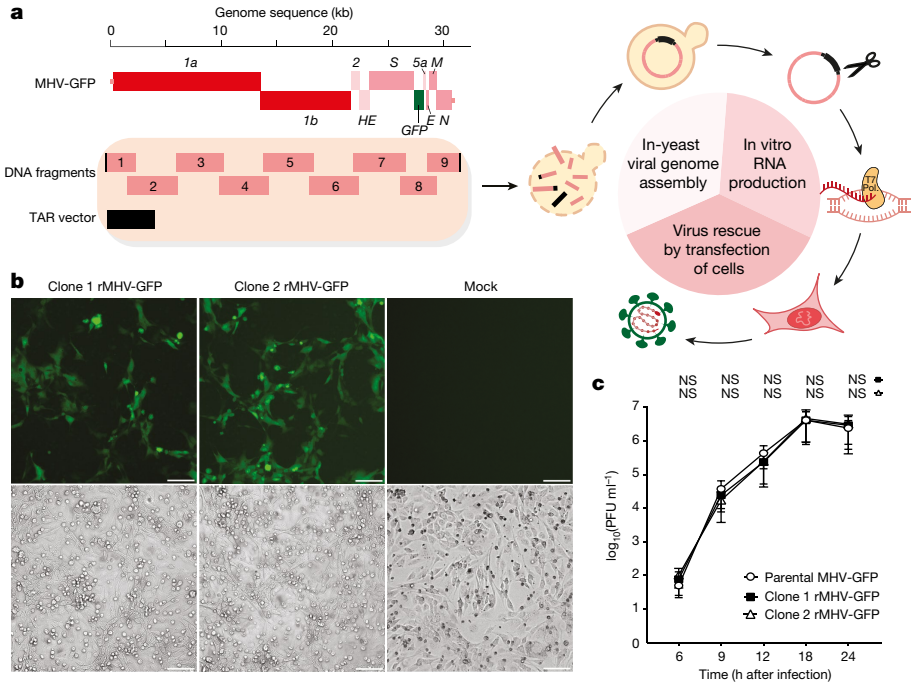


Fig. 1 | Application of yeast-based TAR cloning to generate viral cDNA clones and the recovery of recombinant MHV-GFP. a, General workflow of TAR cloning and virus rescue. In-yeast genome reconstruction requires one-step delivery of overlapping DNA fragments that cover the viral genome and a TAR vector in yeast. Viral ORFs and the ORF for GFP are indicated. Transformed DNA fragments are assembled by homologous recombination in yeast to generate a YAC that contains the full-length viral cDNA sequence. In vitro production of infectious capped viral RNA starts with the isolation of the YAC, followed by plasmid linearization to provide a DNA template for run-off T7 RNA polymerase-based transcription. Virus rescue is initiated by electroporation of BHK-MHV-N cells, after which virus production and amplification is carried out by culturing the virus with susceptible cells. **b**, Recovery of infectious rMHV-GFP from yeast clones 1 and 2. Cell-culture supernatants—which contain viruses produced after virus rescue of two MHV-GFP YAC clones—were used to infect 17Cl-1 cells. At 48 h after infection, infected cells were visualized for GFP expression (top) and by bright-field microscopy (bottom). Mock represents 17Cl-1 cells inoculated with the supernatant from BHK-MHV-N cells electroporated without viral RNAs. Images are representative of two independent experiments. Scale bars, 100 μm. **c**, Replication kinetics of parental MHV-GFP and rMHV-GFP clones 1 and 2. L929 cells were infected (multiplicity of infection (MOI) = 0.1), and cell-culture supernatants were collected at the indicated time points after infection and titrated by plaque assay. PFU, plaque forming units. Data represent the mean ± s.d. of three independent biological experiments ($n = 3$). Statistical significance was determined by two-sided unpaired Student's *t*-test without adjustments for multiple comparisons. NS, not significant. *P* values (from left to right): top, NS, $P = 0.2905$; NS, $P = 0.3504$; NS, $P = 0.1817$; NS, $P = 0.9862$; NS, $P = 0.6738$; bottom, NS, $P = 0.0835$; NS, $P = 0.1400$; NS, $P = 0.2206$; NS, $P = 0.8020$; NS, $P = 0.5894$.

the large genome size (27–31 kb) and occasional instability of cloned DNA in *E. coli*. However, unconventional approaches—such as cloning in low-copy bacterial artificial chromosomes (BACs) or vaccinia virus, or cloning of subgenomic DNA fragments followed by in vitro ligation—were successful[1–3], although each system has caveats that make the generation of recombinant coronavirus genomes cumbersome. Here we assessed the suitability of the yeast *S. cerevisiae* to assemble and maintain genomes of diverse RNA viruses to establish a rapid, stable and universal reverse-genetics pipeline for RNA viruses.

To generate a yeast-based reverse-genetics platform for RNA viruses, we first used mouse hepatitis virus (MHV) strain A59, which contains the gene for green fluorescent protein (MHV-GFP) and which has an established vaccinia virus-based reverse-genetics platform[14,15]. The overall strategy is shown in Fig. 1a. Viral RNA was prepared from MHV-GFP-infected mouse 17Cl-1 cells and used to amplify seven overlapping DNA fragments by reverse-transcription PCR (RT–PCR) that spanned the MHV-GFP genome from nucleotides 2024 to 29672. Fragments containing the 5′ and 3′ termini were PCR-amplified from the vaccinia virus-cloned genome to include a T7 RNA polymerase promoter directly upstream of the MHV-GFP 5′ end and a cleavage site (PacI) after the poly(A) sequence at the MHV-GFP 3′ end, which is required to produce RNA run-off transcripts using T7 RNA polymerase[14]. Overlap sequences for the TAR vector pVC604 were included in the primers that amplified the 5′- and 3′-terminal fragments (Supplementary Table 1). All DNA fragments were simultaneously transformed into *S. cerevisiae* (strain VL6-48N), and the resulting clones were screened for the correct assembly of the yeast artificial chromosome (YAC) containing the cloned MHV genome by multiplex PCRs that covered the junctions between recombined fragments. This screen revealed that more than 90% of the clones tested were positive, indicating that the assembly in yeast is highly efficient (Supplementary Fig. 1a). To rescue MHV-GFP, we randomly chose two clones, purified and linearized the YACs using PacI (Extended Data Table 1) and subjected the YACs to in vitro transcription using T7 RNA polymerase to generate capped viral genomic RNA. This RNA was transfected together with an in vitro-transcribed mRNA that encodes the MHV nucleocapsid (N) protein into BHK-MHV-N cells, which were then mixed with MHV-susceptible 17Cl-1 cells as previously described[14]. Cytopathogenic effects, virus-induced syncytia and GFP-expressing cells were readily detectable for both clones within 48 h, indicating the successful recovery of infectious virus (Fig. 1b). Finally, we assessed the replication kinetics of the recovered viruses, which were indistinguishable from the parental MHV-GFP line (Fig. 1c).

To address whether the synthetic genomics platform can be applied to other coronaviruses and whether it can be used for rapid mutagenesis, we used a molecular BAC clone of MERS-CoV[16]. We PCR-amplified eight overlapping DNA fragments that covered the MERS-CoV genome (Extended Data Fig. 1a, Supplementary Fig. 1b and Supplementary Table 1). The 5′- and 3′-terminal fragments contained the T7 RNA polymerase promoter upstream of the MERS-CoV 5′ end and the restriction endonuclease cleavage site MluI downstream of the poly(A) sequence,

Table 1 | RNA virus genomes cloned using the synthetic genomics platform

Virus	Family	Size (kb)	Template	Fragment generation	Number of fragments	Virus rescue
MHV-GFP	*Coronaviridae*	31.9	Viral RNA, DNA clone	RT–PCR, PCR	9	Yes
MERS-CoV	*Coronaviridae*	30.1	DNA clone	PCR	8	Yes
MERS-CoV-GFP	*Coronaviridae*	30.7	DNA clone, GFP plasmid DNA	PCR	10	Yes
HCoV-229E	*Coronaviridae*	27.3	Viral RNA, DNA clone	RT–PCR, PCR	13	Not attempted
HCoV-HKU1	*Coronaviridae*	29.9	Synthetic DNA, viral RNA	PCR, RT–PCR	11	Not attempted
MERS-CoV Riyadh-1734-2015	*Coronaviridae*	30	Viral RNA	RT–PCR	8	Not attempted
ZIKA virus	*Flaviviridae*	10.8	Viral RNA	RT–PCR	6	Not attempted
Human RSV-B	*Pneumoviridae*	15	Clinical sample	RT–PCR	4	Not attempted
SARS-CoV-2	*Coronaviridae*	30	Synthetic DNA, viral RNA	Plasmid, RT–PCR	12	Yes
SARS-CoV-2-GFP	*Coronaviridae*	30.5	Synthetic DNA, viral RNA	Plasmid, RT–PCR/ PCR	14	Yes
synSARS-CoV-2-GFP	*Coronaviridae*	30.5	Synthetic DNA	Plasmid, PCR	19	Yes

The number of fragments excludes the TAR vector fragment.

and overlapping sequences with the TAR plasmid pVC604. To mutagenize the MERS-CoV clone, fragment 7 was divided into three overlapping PCR fragments to place the *GFP* gene in frame with a porcine teschovirus 2A element and open-reading frame 4a (ORF4a)[16] (Extended Data Fig. 1a and Supplementary Table 1). Again, almost all YAC clones were successfully assembled (Supplementary Fig. 1b, c). Virus rescue from cloned DNA was performed as described previously[16], resulting in recombinant (r)MERS-CoV and rMERS-CoV-GFP (Extended Data Fig. 1b). This demonstrates that the synthetic genomics platform is suitable to genetically modify coronavirus genomes. As expected, the replication kinetics of rMERS-CoV and rMERS-CoV-GFP were slightly reduced compared with the cell-culture-adapted MERS-CoV-EMC strain (Extended Data Fig. 1c).

Next, we thoroughly evaluated the stability of the cloned genomes, the range of applicability to other virus genomes and whether molecular clones can be generated from clinical samples. Yeast clones that contained YACs encoding MHV-GFP and MERS-CoV were passaged 15–17 times, and sequencing revealed that the genomes could be stably maintained (Extended Data Table 2). We further cloned several other coronaviruses (HCoV-229E[2], HCoV-HKU1 (GenBank: NC_006577) and MERS-CoV-Riyadh-1734-2015 (GenBank: MN481979)) and viruses of other families, such as ZIKA virus (family *Flaviviridae*, GenBank: KX377337) and human respiratory syncytial virus (hRSV; family *Pneumoviridae*) (Table 1), which are known to be difficult to clone and stably maintain in *E. coli*. As shown in Supplementary Fig. 1d–h, cloning of these viral genomes in yeast was in all cases successful irrespectively of the virus source, the nucleic acid template or the number of DNA fragments. Of note, we cloned hRSV-B without any prior information on the virus genotype directly from a clinical sample (nasopharyngeal aspirate) by designing RSV consensus primers to amplify four overlapping DNA fragments (Supplementary Table 1) (sequence submitted to GenBank: MT107528). Collectively, these results demonstrate that the synthetic genomics platform provides the technical advance to rapidly generate molecular clones of diverse RNA viruses by using virus isolates, cloned DNA, synthetic DNA or clinical samples as starting material.

The detection of a new coronavirus in China at the end of 2019 prompted us to test the applicability of our synthetic genomics platform to reconstruct the virus based on the genome sequences released on 10–11 January 2020 (Fig. 2). We divided the genome into 12 overlapping DNA fragments (Fig. 3a, Extended Data Table 3, Supplementary Fig. 1i and Supplementary Table 1). In parallel, we aimed to generate a SARS-CoV-2 clone that expressed GFP, as this could facilitate the screening of antiviral compounds and be used to establish diagnostic assays (for example, virus neutralization assays). This was achieved by dividing fragment 11 into three subfragments (Fig. 3a, Supplementary Fig. 1j and Supplementary Table 1), and GFP was inserted in-frame of ORF7a, replacing nucleotides 40–282. We noticed that nucleotides 3–5 at the 5′ end of the reported SARS-CoV-2 sequence (5′-AU**UAA**AGG; GenBank MN996528.1; nucleotides that are different are highlighted in bold) differed from SARS-CoV (5′-AU**AUU**AGG; GenBank AY291315) and from the more closely related bat SARS-related CoVs ZXC21 and ZC45 (5′-AU**AUU**AGG)[4,17,18] (Extended Data Fig. 2a, b). We therefore designed three 5′-end versions, and each version was combined with the remaining SARS-CoV-2 genome (constructs 1–3) or a corresponding SARS-CoV-2-GFP genome (constructs 4–6). Constructs 1 and 4 contained the 5′ end modified by three nucleotides according to the bat SARS-related CoVs (5′-AU**AUU**AGG), constructs 2 and 5 contained the 124 5′-terminal nucleotides of SARS-CoV, and constructs 3 and 6 contained the reported SARS-CoV-2 sequence (5′-AU**UAA**AGG; according to MN996528.1) (Extended Data Fig. 2a, b). Notably, differences between SARS-CoV-2 and SARS-CoV within the 5′-terminal 124 nucleotides are in agreement with the predicted RNA secondary structures (Extended Data Fig. 2b).

Fourteen synthetic DNA fragments were ordered as sequence-confirmed plasmids and all but fragments 5 and 7 were delivered (Extended Data Table 3, Supplementary Data 1). As we received SARS-CoV-2 viral RNA from an isolate of a Munich patient (BetaCoV/Germany/BavPat1/2020) at the same time, we amplified the regions of fragments 5 and 7 by RT–PCR (Supplementary Table 1). TAR cloning was immediately initiated, and for all six SARS-CoV-2 and SARS-CoV-2-GFP constructs we obtained correctly assembled molecular clones (Extended Data Fig. 3a and Supplementary Fig. 1i, j). Because sequence verification was not possible within this short time frame, we randomly selected two clones for each construct (Extended Data Fig. 3a), isolated the YAC DNA and performed in vitro transcription. The resulting RNAs were electroporated together with an mRNA that encodes the SARS-CoV-2 N protein into BHK-21 and, in parallel, into BHK-SARS-N cells that expressed the SARS-CoV N protein[19] (Extended Data Fig. 3b). Electroporated cells were seeded on Vero E6 cells and two days later we observed green fluorescent signals in cells that received the GFP-encoding SARS-CoV-2 RNAs. Indeed, we could rescue infectious viruses for almost all rSARS-CoV-2 and rSARS-CoV-2-GFP clones (Extended Data Fig. 3b). As shown in Fig. 3b, for rSARS-CoV-2 clones

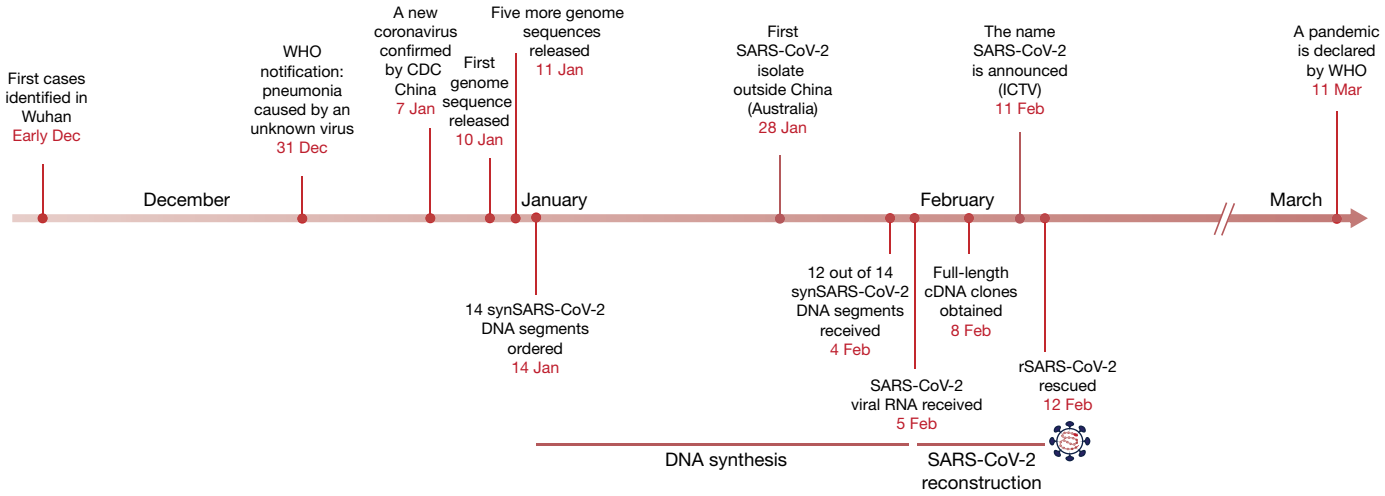


Fig. 2 | Timeline of the reconstruction and recovery of rSARS-CoV-2 in relation to key events of the COVID-19 pandemic. Illustration of the rapidity of rSARS-CoV-2 reconstruction along with the timeline of key events of the COVID-19 pandemic. CDC, Center for Disease Control and Prevention; ICTV, International Committee on Taxonomy of Viruses; WHO, World Health Organization.

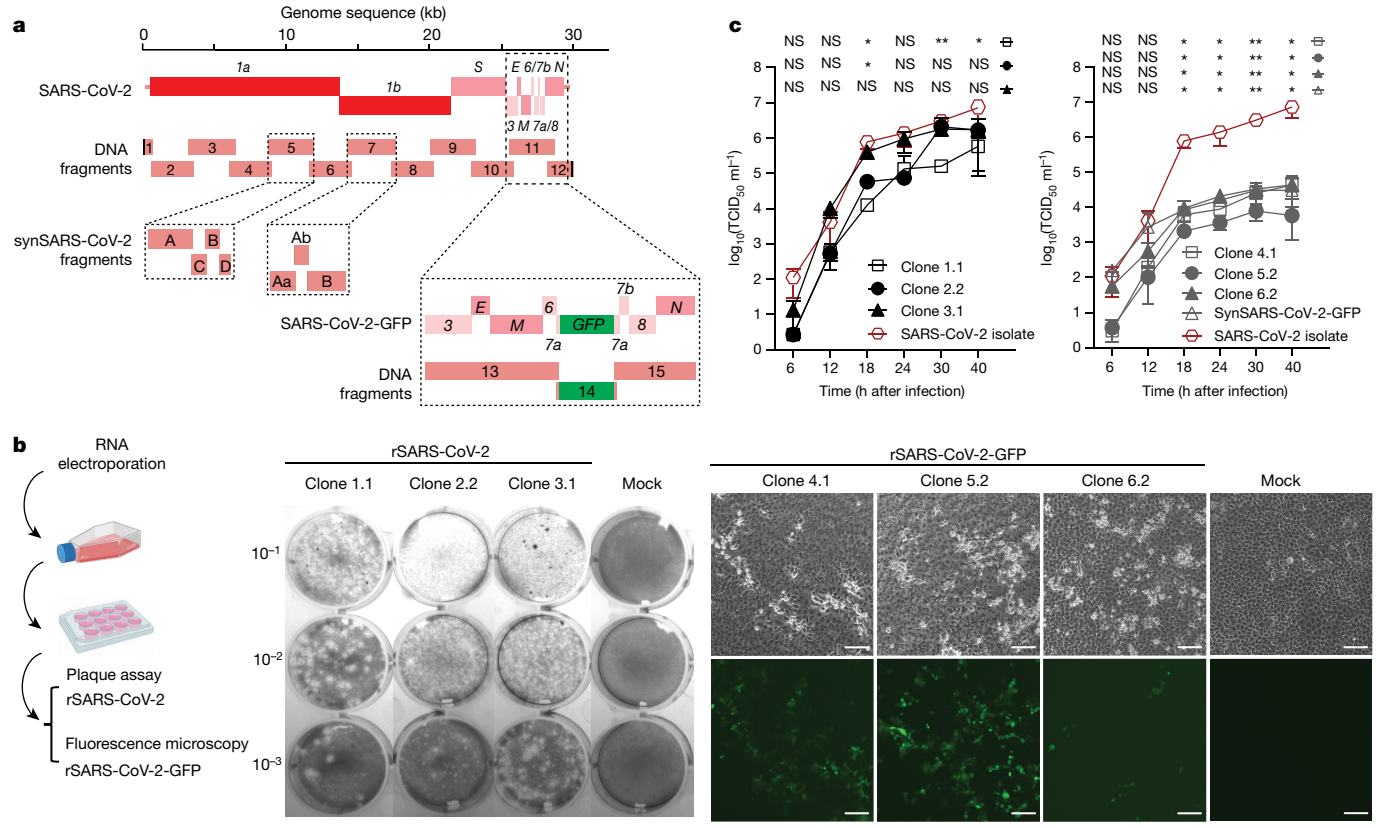


Fig. 3 | Reconstruction, rescue and characterization of rSARS-CoV-2, rSARS-CoV-2-GFP and synSARS-CoV-2-GFP. a, Schematic representation of the SARS-CoV-2 genome organization and DNA fragments used to clone rSARS-CoV-2, rSARS-CoV-2-GFP and synSARS-CoV-2-GFP. Inserts show synthetic subfragments comprising fragments 5 (A–D) and 7 (Aa, Ab, B), and the fragments used to insert the *GFP* gene (fragments 13–15). **b**, Left, schematic of the experiment. Middle, rescue of rSARS-CoV-2 from yeast clones 1.1, 2.2 and 3.1. Supernatants (10^{-1}, 10^{-2} and 10^{-3} ml) of cells infected with the indicated clones or mock-infected cells were transferred to Vero E6 cells to detect plaques (rSARS-CoV-2). Right, rescue of rSARS-CoV-2-GFP from yeast clones 4.1, 5.2 and 6.2. Supernatants (1 ml) from individual rescue experiments were transferred to Vero E6 cells to detect green fluorescence (rSARS-CoV-2-GFP). Mock, uninfected cells. Scale bars, 100 µm. **c**, Replication kinetics of rSARS-CoV-2 clones 1.1, 2.2, 3.1 (left) and rSARS-CoV-2-GFP clones 4.1, 5.2, 6.2 and synSARS-CoV-2-GFP (right) compared with the SARS-CoV-2 isolate. Vero E6 cells were infected (MOI = 0.01), and supernatants were collected at the indicated time points after infection and titrated (50% tissue culture infectious dose ($TCID_{50}$) assay). Data represent the mean ± s.d. of three independent biological replicates. Statistical significance was determined for each clone against the SARS-CoV-2 isolate by two-sided unpaired Student's *t*-test without adjustments for multiple comparisons. *P* values (from left to right): left, top, NS, $P = 0.0851$; NS, $P = 0.1775$; $^{*}P = 0.0107$; NS, $P = 0.0648$; $^{**}P = 0.0013$; $^{*}P = 0.0373$; middle, NS, $P = 0.0851$; NS, $P = 0.1713$; $^{*}P = 0.0133$; NS, $P = 0.0535$; NS, $P = 0.0909$; NS, $P = 0.0632$; bottom, NS, $P = 0.1119$; NS, $P = 0.1641$; NS, $P = 0.0994$; NS, $P = 0.4921$; NS, $P = 0.3336$; NS, $P = 0.0790$; right, top, NS, $P = 0.0858$; NS, $P = 0.1429$; $^{*}P = 0.0104$; $^{*}P = 0.0466$; $^{**}P = 0.0011$; $^{*}P = 0.0287$; second, NS, $P = 0.0872$; NS, $P = 1360$; $^{*}P = 0.0102$; $^{*}P = 0.0461$; $^{**}P = 0.0011$; $^{*}P = 0.0282$; third, NS, $P = 0.4810$; NS, $P = 0.1758$; $^{*}P = 0.0106$; $^{*}P = 0.0478$; $^{**}P = 0.0011$; $^{*}P = 0.0287$; bottom, NS, $P = 0.3739$; NS, $P = 0.6817$; $^{*}P = 0.0106$; $^{*}P = 0.0473$; $^{**}P = 0.0011$ $^{*}P = 0.0285$.

1.1, 2.2, and 3.1, plaques were readily detectable, demonstrating that infectious virus has been recovered irrespectively of the 5′-terminal sequences. Sequencing of the YACs and corresponding rescued viruses revealed that almost all DNA clones and viruses contained the correct sequence, except for some individual clones that contained mutations within fragments 5 and 7 that were probably introduced by RT–PCR (Extended Data Table 4). Nevertheless, we obtained at least one correct YAC clone for all constructs except for construct 6. To correct this, we reassembled construct 6 by replacing the RT–PCR-generated fragments 5 and 7 with four and three shorter synthetic double-stranded (ds)DNA fragments, respectively. The resulting molecular clone was used to rescue the synthetic SARS-CoV-2-GFP (synSARS-CoV-2-GFP) virus without any mutations exclusively from chemically synthesized DNA (Extended Data Fig. 4 and Extended Data Tables 3, 4).

Next we assessed the 5′ end of the recombinant viruses and the Munich virus isolate and confirmed the published 5′ end sequence of SARS-CoV-2 (5′-AU**UAA**AGG; GenBank MN996528.3). Full-length sequencing of the viral genomes and 5′ rapid amplification of cDNA end (5′-RACE) analysis of the recombinant viruses confirmed the identity of each virus, and showed that the 5′ end variant of each virus retained the cloned 5′ terminus (Extended Data Fig. 2a). This demonstrates that the 5′ ends of SARS-CoV and bat SARS-related CoVs ZXC21 and ZC45 are compatible with the replication machinery of SARS-CoV-2. Sequencing results also revealed the identity of leader–body junctions of SARS-CoV-2 subgenomic mRNAs, which are identical to those of SARS-CoV[18] (Extended Data Fig. 2c–h). We also analysed rSARS-CoV-2 clone 3.1 for protein expression and demonstrated the presence of the SARS-CoV-2 nucleocapsid protein in dsRNA-positive cells (Extended Data Fig. 5b). The replication kinetics of rSARS-CoV-2 clone 3.1, which contains the authentic 5′ terminus, was indistinguishable from replication of the SARS-CoV-2 isolate, while clones 1.1 and 2.2 showed slightly reduced replication (Fig. 3c, left). All rSARS-CoV-GFP clones and synSARS-CoV-GFP displayed similar growth kinetics but they were significantly reduced compared with the SARS-CoV-2 isolate, suggesting that the insertion of GFP and/or the partial deletion of ORF7a affects replication (Fig. 3c, right and Extended Data Fig. 5d–f). Despite the reduced replication, green fluorescence was readily detectable and we demonstrated the use of the synSARS-CoV-GFP clone for antiviral drug screening by testing remdesivir, a promising compound for the treatment of COVID-19[20] (Extended Data Fig. 5c). Similarly, the simple readout of green fluorescence greatly facilitates the demonstration of virus neutralization with human serum (Extended Data Fig. 5a).

Our results demonstrate the full functionality of the SARS-CoV-2 reverse-genetics system and we expect that this fast, robust and versatile synthetic genomics platform will provide new insights into the molecular biology and pathogenesis of a number of emerging RNA viruses. Although homologous recombination in yeast has already been used for the generation of a number of molecular virus clones in the past[12,13,21,22], we present a thorough evaluation of the feasibility of this approach to rapidly generate full-length cDNAs for large RNA viruses that have a known history of instability in *E. coli*. We show that one main advantage of the TAR cloning system is that the viral genomes can be fragmented to at least 19 overlapping fragments and reassembled with remarkable efficacy. This facilitated the cloning and rescue of rSARS-CoV-2 and rSARS-CoV-2-GFP within one week. It should be noted that we see considerable potential to reduce the time of DNA synthesis. Currently, synthetic DNA fragments get routinely cloned in *E. coli*, which turned out to be problematic for SARS-CoV-2 fragments 5 and 7. We, however, used shorter synthetic dsDNA parts to assemble these fragments by TAR cloning and to generate the molecular clone synSARS-CoV-2-GFP by using exclusively chemically synthesized DNA, which is an additional proof of the superior cloning efficiency of yeast- versus *E. coli*-based systems.

The COVID-19 pandemic emphasizes the need for preparedness to rapidly respond to emerging virus threats. The rapidity of our synthetic genomics approach to generate SARS-CoV-2 and the applicability to other emerging RNA viruses make this system an attractive alternative to provide infectious virus samples to health authorities and diagnostic laboratories without the need of having access to clinical samples. As the COVID-19 pandemic is ongoing, we expect to see sequence variations and possibly phenotypic changes of the evolving SARS-CoV-2 virus in the human host. With this synthetic genomics platform, it is now possible to rapidly introduce such sequence variations into the infectious clone and to functionally characterize SARS-CoV-2 evolution in real time.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2294-9.

1. Almazán, F. et al. Engineering the largest RNA virus genome as an infectious bacterial artificial chromosome. *Proc. Natl Acad. Sci. USA* **97**, 5516–5521 (2000).
2. Thiel, V., Herold, J., Schelle, B. & Siddell, S. G. Infectious RNA transcribed in vitro from a cDNA copy of the human coronavirus genome cloned in vaccinia virus. *J. Gen. Virol.* **82**, 1273–1281 (2001).
3. Yount, B., Curtis, K. M. & Baric, R. S. Strategy for systematic assembly of large RNA and DNA genomes: transmissible gastroenteritis virus model. *J. Virol.* **74**, 10600–10611 (2000).
4. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
5. Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. & Fouchier, R. A. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N. Engl. J. Med.* **367**, 1814–1820 (2012).
6. Cao-Lormeau, V. M. et al. Zika virus, French Polynesia, South Pacific, 2013. *Emerg. Infect. Dis.* **20**, 1084–1086 (2014).
7. Baize, S. et al. Emergence of Zaire Ebola virus disease in Guinea. *N. Engl. J. Med.* **371**, 1418–1425 (2014).
8. Gibson, D. G. et al. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).
9. Lartigue, C. et al. Creating bacterial strains from genomes that have been cloned and engineered in yeast. *Science* **325**, 1693–1696 (2009).
10. Benders, G. A. et al. Cloning whole bacterial genomes in yeast. *Nucleic Acids Res.* **38**, 2558–2569 (2010).
11. Kouprina, N. & Larionov, V. Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast *Saccharomyces cerevisiae*. *Nat. Protoc.* **3**, 371–377 (2008).
12. Vashee, S. et al. Cloning, assembly, and modification of the primary human cytomegalovirus isolate Toledo by yeast-based transformation-associated recombination. *mSphere* **2**, e00331-17 (2017).
13. Oldfield, L. M. et al. Genome-wide engineering of an infectious clone of herpes simplex virus type 1 using synthetic genomics assembly methods. *Proc. Natl Acad. Sci. USA* **114**, E8885–E8894 (2017).
14. Coley, S. E. et al. Recombinant mouse hepatitis virus strain A59 from cloned, full-length cDNA replicates to high titers in vitro and is fully pathogenic in vivo. *J. Virol.* **79**, 3097–3106 (2005).
15. Züst, R. et al. Coronavirus non-structural protein 1 is a major pathogenicity factor: implications for the rational design of coronavirus vaccines. *PLoS Pathog.* **3**, e109 (2007).
16. Muth, D. et al. Transgene expression in the genome of Middle East respiratory syndrome coronavirus based on a novel reverse genetics system utilizing Red-mediated recombination cloning. *J. Gen. Virol.* **98**, 2461–2469 (2017).
17. Hu, D. et al. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. *Emerg. Microbes Infect.* **7**, 1–10 (2018).
18. Thiel, V. et al. Mechanisms and enzymes involved in SARS coronavirus genome expression. *J. Gen. Virol.* **84**, 2305–2315 (2003).
19. van den Worm, S. H. et al. Reverse genetics of SARS-related coronavirus using vaccinia virus-based recombination. *PLoS ONE* **7**, e32857 (2012).
20. Sheahan, T. P. et al. Comparative therapeutic efficacy of remdesivir and combination lopinavir, ritonavir, and interferon beta against MERS-CoV. *Nat. Commun.* **11**, 222 (2020).
21. Polo, S., Ketner, G., Levis, R. & Falgout, B. Infectious RNA transcripts from full-length dengue virus type 2 cDNA clones made in yeast. *J. Virol.* **71**, 5366–5374 (1997).
22. Nikiforuk, A. M. et al. Rapid one-step construction of a Middle East respiratory syndrome (MERS-CoV) infectious clone system by homologous recombination. *J. Virol. Methods* **236**, 178–183 (2016).

Methods

Cells and general culture conditions

Vero, Vero B4 and Vero B6 cells (all ATCC) were cultured in Dulbecco's modified Eagle's medium (DMEM); BHK-21, BHK-MHV-N (BHK-21 cells expressing the N protein of MHV strain A59)[14], BHK-SARS-N (BHK-21 cells expressing the N protein of SARS)[19], Huh-7[23], L929[23] and mouse 17Cl-1[23] cells were grown in minimal essential medium (MEM). Both types of medium were supplemented with 10% fetal bovine serum, 1× non-essential amino acids, 100 units ml^{-1} penicillin and 100 μg ml^{-1} streptomycin. BHK-SARS-N cells were grown using MEM supplemented with 5% fetal bovine serum, 1× non-essential amino acids, 100 units ml^{-1} penicillin, and 100 μg ml^{-1} streptomycin, 500 μg ml^{-1} G418 and 10 μg ml^{-1} puromycin. BHK-MHV-N and BHK-SARS-N were treated with 1 μg ml^{-1} doxycyclin 24 h before electroporation. All cells were maintained at 37 °C and in a 5% CO_2 atmosphere.

Cultured viruses

MHV-GFP[14,15] and HCoV-229E[2] were cultured in mouse 17Cl-1 and human Huh-7 cells, respectively. MERS-CoV-EMC[24] was cultured in Vero B4 cells. HCoV-HKU1 strain Caen-1 (GenBank: NC_006577) was cultured in human airway epithelial cultures[25]. ZIKA virus strain PRVABC-59 (GenBank: KX377337) was provided by M. Alves and was cultured in Vero cells. SARS-CoV-2 (SARS-CoV-2/München-1.1/2020/929) was cultured in Vero E6 cells.

Bacterial and yeast strains

E. coli DH5α (Thermo Scientific) and TransforMax Epi300 (Epicentre) were used to propagate the pVC604 and pCC1BAC-His3 TAR vectors[8], respectively. The bacteria were grown in lysogeny broth medium supplemented with the appropriate antibiotics at 37 °C overnight. *E. coli* Epi300 cells containing the different synthetic fragments of SARS-CoV-2 in pUC57 or pUC57mini were grown at 30 °C to decrease the risk of instability and/or toxicity. *Saccharomyces cerevisiae* VL6-48N (MATα trp1-Δ1 ura3-Δ1 ade2-101 his3-Δ200 lys2 met14 cir°) was used for all yeast transformation experiments[26]. Yeast cells were first grown in YPDA broth (Takara Bio), and transformed cells were plated on minimal synthetic defined (SD) agar without histidine (SD−His) (Takara Bio). *S. cerevisiae* VL6-48N-derived clones carrying different YACs were never streaked out together on the same agar dishes as mating switching and resulting recombination might occur at a very low frequency.

Generation of viral subgenomic fragments for TAR cloning using viral RNA, infectious cDNA clones and synthetic DNA

Table 1 displays the templates used to clone the different viral genomes into *S. cerevisiae*. In general, viral DNA fragments were obtained by RT–PCR of viral RNA extracted from viral strains, isolates and from clinical specimens, using the SuperScript IV One-Step RT–PCR System following the manufacturer's instructions. Additionally, some fragments were PCR-amplified from vaccinia virus-cloned cDNA[2,14], BAC-cloned cDNA[16] and plasmid-cloned synthetic DNA (GenScript), using the CloneAmp HiFi PCR Premix according to the manufacturer's instructions. Accessory sequences, that is, enhanced GFP and porcine teschovirus-1 2A (P2A) for the MERS-CoV-GFP construct, TurboGFP for SARS-CoV-2-GFP and T7 RNA polymerase promoter-hammerhead ribozyme and ribozyme-T7 terminator for human RSV-B, were amplified from plasmids.

For all coronaviruses, the fragment encompassing the viral 5′ untranslated regions (UTR) contained the T7 RNA polymerase promoter sequence immediately upstream of the 5′ end of the genome, and the fragment encompassing the 3′ end of the genome contained a unique restriction site (Extended Data Table 1) downstream of the poly(A) tail.

HCoV-HKU1 synthetic fragments 1–4 were provided individually cloned into pUC57 by GenScript (Extended Data Table 3). MERS-CoV-Riyadh-1734-2015 (GenBank: MN481979) fragments 1–8 were synthesized and cloned into pUC57 by GenScript (Extended Data Table 3), containing homologous regions to TAR vectors pVC604 and pCC1BAC-His3. Similarly, synthetic ZIKA virus fragment 6 cloned into pUC57 contained a hepatitis delta virus ribozyme sequence and pCC1BAC-his3 homology downstream of the viral 3′ UTR (Extended Data Table 3).

The SARS-CoV-2 synthetic DNA fragments were delivered cloned into pUC57 or pUC57mini by GenScript (Supplementary Data 1, Extended Data Table 3). Fragments 1.1, 1.2, 1.3 and 12 contained homologous sequences to pCC1BAC-His3. Each fragment was sequence verified using Sanger sequencing after plasmid isolation using QIAGEN Midiprep kit (QIAGEN). Fragments were released from the vector using the restriction enzymes described in Extended Data Table 3. Restricted fragments were subsequently gel-purified using standard methods[27]. DNA concentrations and purities of all fragments to be used for TAR cloning were determined using NanoDrop 2000/2000c Spectrophotometer (Thermo Scientific).

In-yeast cloning of viral genomes using TAR

In general, we used overlapping DNA fragments for TAR cloning with overlaps ranging from 45 to 500 bp. As all of our cloning experiments worked well, we did not assess whether the lengths of the overlap affected homologous recombination efficacy. The vectors pVC604[11] and pCC1BAC-His3[8] were used for TAR cloning. These vectors were amplified by PCR using primers containing at least 45-bp overlaps to fragments encompassing the 5′ or 3′ ends of different viral genomes (Supplementary Table 1). Amplification was performed using KOD Hot Start DNA polymerase (Merck Millipore) according to the manufacturer's instructions. Templates used for generating fragments for TAR cloning are shown in Table 1. TAR cloning was also used to reconstruct the full-length synthetic fragments 5 and 7 in yeast (Extended Data Fig. 4b, c).

Yeast transformation was done using the high-efficiency lithium acetate/SS carrier DNA/PEG method as described elsewhere[28]. In brief, yeast cells were grown in rich YPDA medium (Takara Bio) at 30 °C with agitation until an optical density at 600 nm of 1.0 was reached. Then, 3 ml of yeast culture was used per transformation event. DNA mixtures were prepared beforehand and contained 100–200 fmol of 3′ and 5′ open ends for all fragments. Transformation mixtures were plated onto SD−His plates (Takara Bio) and incubated at 30 °C for 48 h. Colonies were resuspended in 20 μl of SD−His broth, and DNA was extracted following the GC prep method[29]. Extracted DNA was used as template for screening by multiplex PCR using the QIAGEN Multiplex PCR kit (QIAGEN) according to the manufacturer's instruction. One or two multiplex PCRs were designed to encompass different subsets of primer pairs, and cover all desired recombination junctions (Supplementary Table 1). Clones tested positive for all junctions were grown in SD−His until late logarithmic phase, and plasmids were extracted from 500 ml culture using the QIAGEN Maxiprep Kit (QIAGEN) with modifications. In brief, 10 ml of Buffer P1 was supplemented with 1 ml of zymolyase solution (10 mg ml^{-1} Zymolyase 100-T; 50 mM Tris-HCl pH 7.5; 50% (v/v) glycerol) and 100 μl of β-mercapthoethanol. The mixture was incubated for 1 h at 37 °C before the addition of buffer P2. The rest of the protocol followed the manufacturer's instructions. DNA preparations were successfully used as templates to generate in vitro transcribed viral RNA even if they contained traces of yeast genomic DNA. In parallel, isolated YACs containing full-length synthetic fragments 5 and 7, as well as SARS-CoV-2 and SARS-CoV-2-GFP viral genomes, were successfully transformed into *E. coli* TransforMax Epi300 electrocompetent cells (Epicentre) (data not shown).

Stability testing of the YAC containing entire RNA virus genomes in yeast

The stability of viral genomes maintained as YACs in *S. cerevisiae* was tested for the clones containing MHV-GFP or MERS-CoV for 1 week.

A single colony was grown in 20 ml of SD−His liquid medium, 1 ml aliquots were removed and expanded in fresh medium every 12 h. The generation time for each of the clones was estimated to range from 150 to 160 min. After 15–17 passages, each YAC clone was isolated and subjected to sequencing by MinION (Oxford Nanopore Technologies) to obtain the entire YAC sequence. Individual regions for which MinION sequencing did not reveal a clear sequence were resequenced by Sanger sequencing (Microsynth).

Virus rescue

The YAC containing viral cDNA was cleaved at the unique restriction site located downstream of the 3′ end poly(A) tail (Extended Data Table 1). In brief, 1–2 μg of phenol–chloroform-extracted and ethanol-precipitated restricted DNA was resolved in nuclease-free water and used for in vitro transcription using the T7 RiboMAX Large Scale RNA production system (Promega) with m7G(5′)ppp(5′)G cap provided as described previously[2]. Additionally, a similar protocol was performed on a PCR product of the *N* gene from corresponding coronaviruses, producing a capped mRNA that encodes the N protein. Then, 1–10 μg of in vitro transcribed viral RNA was electroporated together with 2 μg of the *N* gene transcript into BHK-21 cells and/or BHK-21 cells expressing the corresponding coronavirus N protein. Electroporated cells were co-cultured with susceptible mouse 17Cl-1, Vero B4 and Vero E6 cells to rescue rMHV-GFP (17Cl-1), rMERS-CoV and rMERS-CoV-GFP (Vero B4), and rSARS-CoV-2, rSARS-CoV-2-GFP and synSARS-CoV-2-GFP (VeroE6). Progeny viruses that were collected from the supernatant immediately after electroporation were termed passage 0 viruses and were used to produce stocks for subsequent analysis. Virus-infected cells were monitored, and images were acquired using an EVOS fluorescence microscope equipped with a 10× air objective. Brightness and contrast were adjusted using FIJI. Figures were assembled using the FigureJ plugin[30].

All work involving the rescue and characterization of recombinant MERS-CoV, SARS-CoV and SARS-CoV-2 was performed in a biosafety level 3 laboratory at the Institute of Virology and Immunology, Mittelhäusern, Switzerland under appropriate safety measures with respect to personal and environmental protection.

Virus growth kinetics

In brief, 24 h before infection with MHV-GFP, L929 cells were seeded in a 24-well plate at a density of 3.6×10^5 cells per ml. Cells were washed once with PBS and inoculated with viruses (multiplicity of infection (MOI) = 0.1). After 2 h, the virus-containing supernatant was removed, and cells were washed three times with PBS and supplied with medium as described above. Cell-culture supernatants were collected at the indicated time points after infection. A similar protocol was used for MERS-CoV and MERS-CoV-GFP using Vero B4 cells (MOI = 0.01), and SARS-CoV-2 using Vero E6 cells (MOI = 0.01). Statistical significance was determined by two-sided unpaired Student's *t*-test without adjustments for multiple comparisons.

Plaque assay and $TCID_{50}$

MHV-GFP PFU ml^{-1} was determined by plaque assay in L929 cells as described previously[14]. In brief, 24 h before infection, L929 cells were seeded in a 24-well plate at a density of 3.6×10^5 cells per ml. Cells were washed with PBS and inoculated with viruses serially diluted in cell-culture medium at 1:10 dilution. Cells were washed with PBS 1 h after inoculation, and overlaid with 2% methylcellulose mixed at 1:1 with 2× DMEM supplemented with 20% fetal bovine serum, 200 units ml^{-1} penicillin and 200 μg ml^{-1} streptomycin. After 24 h of incubation, the overlay was removed and cells were fixed and stained with crystal violet.

The $TCID_{50}$ assay was performed for MERS-CoV and MERS-CoV-GFP in Vero B4 cells and SARS-CoV-2 and SARS-CoV-2-GFP in Vero E6 cells. In brief, cells were seeded 24 h before infection in a 96-well plate at a density of 2×10^6 cells per plate. Viruses were serially diluted at 1:10 dilution from 10^{-1} to 10^{-8}. After 72 h of incubation, the medium was removed and cells were fixed and stained with crystal violet. The $TCID_{50}$ ml^{-1} titre was determined using the Spearman–Kaerber method[31].

The PFU ml^{-1} of SARS-CoV-2 and SARS-CoV-2-GFP was determined by plaque assay using Vero E6 cells in a 6-well format. In brief, 24 h before infection, Vero E6 cells were seeded at a density of 2×10^6 cells per plate. At the time of infection, cells were washed with PBS and inoculated with viruses serially diluted in cell-culture medium at 1:10 dilution. Cells were washed with PBS 1 h after inoculation and overlaid with 2.4% Avicel mixed at 1:1 with 2× DMEM supplemented with 20% fetal bovine serum, 200 units ml^{-1} penicillin and 200 μg ml^{-1} streptomycin. After 48 h of incubation, the overlay was removed and cells were fixed and stained with crystal violet.

Sequencing and computational analysis

Full-length sequences of the SARS-CoV-2 and SARS-CoV-2-GFP cDNAs cloned in yeast were confirmed by Sanger sequencing (Microsynth). All other virus genomes cloned in yeast were confirmed using the Nanopore sequencer MinION from Oxford Nanopore Technologies according to standard protocols. The operating software MinKNOW performed data acquisition and real-time base calling, generating data as fast5 and/or fastq files. Subsequently, the Python command line qcat (Mozilla Public License 2.0., copyright 2018 Oxford Nanopore Technologies, v1.1.0, http://www.github.com/nanoporetech/qcat) was run to demultiplex Nanopore reads from fastq files. Alignment of demultiplexed reads to reference sequences was carried out using the Minimap2 program[32], producing a fasta file. Mutations of consensus sequences and regions for which the sequences were not clear were verified by Sanger sequencing (Microsynth).

rSARS-CoV-2 and SARS-CoV-2-GFP RNA was sequenced by next-generation sequencing using poly(A)-purified RNA. In brief, 1×10^6 Vero E6 cells were infected with rSARS-CoV-2 clones 1.1, 2.2, 3.1 and rSARS-CoV-2-GFP clones 4.1, 5.2, 6.2 (all passage 1) at an MOI = 0.001. Cellular RNA was prepared using NucleoSpin RNA Plus (Macherey-Nagel) according to the manufacturer's recommendation. The quantity and quality of the extracted RNA was assessed using a Thermo Fisher Scientific Qubit 4.0 fluorometer with the Qubit RNA BR Assay Kit (Thermo Fisher Scientific, Q10211) and an Advanced Analytical Fragment Analyzer System using a Fragment Analyzer RNA Kit (Agilent, DNF-471), respectively. Sequencing libraries were produced using an Illumina TruSeq Stranded mRNA Library Prep kit (Illumina, 20020595) in combination with TruSeq RNA UD Indexes (Illumina, 20022371) according to Illumina's guidelines. Pooled cDNA libraries were paired-end sequenced using an Illumina NovaSeq 6000 S Prime Reagent Kit (300 cycles; Illumina, 20027465) on an Illumina NovaSeq 6000 instrument, generating an average of 69 million reads per sample. The quality-control assessments, generation of libraries and sequencing run were all performed at the Next Generation Sequencing Platform, University of Bern, Switzerland. For analysis, the adaptor sequences were trimmed using TrimGalore software (v.0.6.5) and reads shorter than 20 nucleotides in length and/or with a Phred score of less than 20 were removed. Paired-end trimmed reads were mapped to the SARS-CoV-2 genome (GenBank accession MT108784; synthetic construct derived from SARS-2 BetaCoV/Wuhan/IVDC-HB-01/2019) using the Spliced Transcripts Alignment to a Reference (STAR) aligner (v.2.7.0a)[33] with default parameters. Before mapping, STAR was also used to generate a genome index for SARS-CoV-2 with the parameters --genomeSAindexNbases 7 and --sjdbOverhang 149. SAMtools (v.1.10) was used to calculate mapped read depth from the resulting mapped read pairs at each position in the genome and subsequently visualized using a variety of software packages in R. Calculations were performed on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern. Sequencing data have been deposited in the Sequence Read Archive (SRA) of the NCBI (http://www.ncbi.nlm.nih.gov/sra).

Apart from MinION and next-generation sequencing data handling, other sequence analyses were performed using Geneious Prime

v.2019.2.3. Results from virus growth kinetics were analysed and graphically presented using GraphPad Prism v.8.3.0 for Windows. All figures were created with Adobe Illustrator and Biorender.com.

Identification of leader–body junctions of viral mRNAs

To identify reads that mapped discontinuously to the SARS-CoV-2 genome and determine the location of potential transcription regulatory sites (TRS), we pooled reads that mapped to the viral genome as well as unmapped reads and searched for the sequence TTCTCTAA**ACGAAC** (nucleotides 62–75 of MT108784; leader TRS is indicated in bold). We then filtered for reads that had at least 18 nucleotides 3′ of the aforementioned sequence and evaluated whether these reads were compatible with any of the SARS-CoV-2 mRNA sequences. Reads matching these criteria were used as input for the generation of a consensus sequence for each TRS site and analysed using a combination of SAMtools (v.1.10), R and the Integrative Genomics Viewer (IGV). Mapped read depth was also calculated for the discontinuously mapped reads as explained in the previous section.

5′-RACE

Recombinant SARS-CoV-2 and SARS-CoV-2-GFP poly(A)-purified RNA used for next-generation sequencing was also used to determine the genome 5′ ends by 5′-RACE. M-MLV reverse transcription (Promega) was performed according to the manufacturer's instructions using the gene-specific primer pWhSF-ORF1a-R18-655 (Supplementary Table 1) and 10 U RNase Inhibitor RNasin plus (Promega) per 25 µl reaction volume. Following reverse transcription, 1 µl RNase H (5 U μl^{-1}, New England Biolabs) per 25 µl reaction was added, and the mixture was incubated at 37 °C for 20 min. The cDNA was immediately purified with the High Pure PCR product purification kit (Roche) according to the manufacturer's instructions. A poly(A) tail was added to the cDNA with Terminal Transferase (New England Biolabs) according to the manufacturer's instructions. Subsequently, a PCR reaction with the tailed cDNA was performed with the primer pair pWhSF-ORF1a-R18-655 and TagRACE_dT16 (Supplementary Table 1) using the HotStarTaq Master Mix (QIAGEN) according to the manufacturer's instructions with a touchdown cycling protocol: 95 °C for 15 min; 15 cycles of 94 °C for 30 s, 65 °C touchdown to 50 °C for 1 min, 72 °C for 1 min; 25 cycles of 94 °C for 30 s, 50 °C for 1 min, 72 °C for 1 min. Subsequently, 1 µl of this reaction was used for a nested re-amplification with the primer pair pWhSF-5utr-R17-273 and TagRACE (Supplementary Table 1) in a final volume of 50 µl following the same cycling protocol as described above. The PCR fragment was purified using the NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) according to the manufacturer's instructions, and the purified PCR fragment was sent to Microsynth for Sanger sequencing with the primer pWhSF-5utr-R17-273 (Supplementary Table 1). Sequencing raw data were assessed using the SeqManTM II sequence analysis software (DNASTAR).

Remdesivir experiment

Remdesivir (MedChemExpress) was dissolved in DMSO and stored at −80 °C in 20 mM stock aliquots. One day before the experiment, Vero E6 cells were seeded in 24-well plates at a density of 8×10^4 cells per well. Cells were infected with synSARS-CoV-2-GFP (passage 1) at MOI = 0.01 or mock-infected as control. Innocula were removed at 1 h after infection, and replaced with medium containing remdesivir (0.2 µM or 2 µM) or the equivalent amount of DMSO. At 48 h after infection, cells were washed once with PBS and incubated in fresh PBS. Images were acquired using an EVOS fluorescence microscope equipped with a 10× air objective. Brightness and contrast were adjusted identically for each condition and their corresponding control using FIJI. Figures were assembled using the FigureJ plugin[30].

Immunofluorescence assay

One day before infection, Vero E6 cells were seeded in a 12-well removable chamber glass slide (Ibidi) at a density of 4×10^4 cells per well. Cells were infected with rSARS-CoV-2 clone 3.1 (passage 2) or mock-infected as control. At 6 and 24 h after infection, cells were washed twice with PBS and fixed with 4% (v/v) neutral-buffered formalin. Cells were washed twice with PBS before permeabilization with 0.1% Triton X-100 and blocking with PBS supplemented with 50 mM NH_4Cl, 0.1% (w/v) saponin and 2% (w/v) BSA (confocal buffer) for 60 min. Primary antibodies (anti-dsRNA, J2, English and Scientific Consulting, 10010500; and anti-SARS-CoV Nucleocapsid (N), Rockland, 200-401-50) and secondary antibodies (donkey anti-rabbit 594, Jackson ImmunoResearch 711-585-152; and donkey anti-mouse 488, Jackson ImmunoResearch 715-545-150) were diluted in confocal buffer. Slides were covered with 0.17-mm thick, high-performance (1.5H) glass coverslips and mounted using ProLong Diamond Antifade mountant containing 4′,6-diamidino-2-phenylindole (DAPI) (Thermo Fisher Scientific). Images were acquired using an EVOS FL Auto 2 Imaging System equipped with a coverslip-correct 40× air objective. Brightness and contrast were adjusted identically for each condition and their corresponding control using FIJI. Figures were assembled using the FigureJ plugin[30].

Serum neutralization assay

One day before the experiment, Vero E6 cells were seeded in a 96-well clear-bottom, black plate at a density of 2×10^6 cells per well. Serum 2 has been described in another study[34] as patient serum ID7 (convalescent human anti-SARS-CoV-2 serum). Serum 4 has been described previously as patient serum CSS 2 (convalescent human anti-SARS-CoV serum)[35]. Sera 1 and 3 were control sera. In brief, all sera were inactivated for 30 min at 56 °C and diluted at 1:10 in OptiMEM. A twofold serial dilution was performed in OptiMEM in a final volume of 50 µl in a separate 96-well plate (dilutions 1:10 to 1:1,280). Then, 50 µl of synSARS-CoV-2-GFP containing 250 $TCID_{50}$ was added to the diluted sera. The serum–virus mixture was incubated at 37 °C for 60 min, and subsequently added to Vero E6 cells. After 1 h of incubation, supernatants were removed and replaced with medium as described above. At 48 h after infection, expression of GFP and cytopathogenic effects were monitored, and images were acquired using an EVOS fluorescence microscope equipped with a 10× air objective. Brightness and contrast were adjusted identically for each condition and their corresponding control using FIJI. Figures were assembled using the FigureJ plugin[30].

Ethical statement

The authors are aware that this work contains aspects of Dual Use Research of Concern (DURC). The benefits were carefully balanced against the risks and the benefits outweigh the risks. Permission to generate and work with recombinant SARS-CoV-2 and SARS-CoV-2-GFP was granted by the Swiss Federal Office of Public Health (A131191/3) with consultation of the Federal Office for Environment, Federal Food Safety and Veterinary Office, and the Swiss Expert Committee for Biosafety.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The following genome sequences are available from GenBank: rSARS-CoV-2 (MT108784), hRSV/B/Bern/2019 (MT107528); MERS-CoV Riyadh-1734-2015 (MN481979). The RNA-sequencing data of rSARS-CoV-2 and rSARS-CoV-2-GFP are available from the NCBI Sequence Read Archive (BioProject accession number PRJNA615319; BioSample accessions: SAMN14450686, SAMN14450687, SAMN14450688, SAMN14450689, SAMN14450690 and SAMN14450691). Source data are provided with this paper.

23. V'kovski, P. et al. Determination of host proteins composing the microenvironment of coronavirus replicase complexes by proximity-labeling. *eLife* **8**, e42037 (2019).

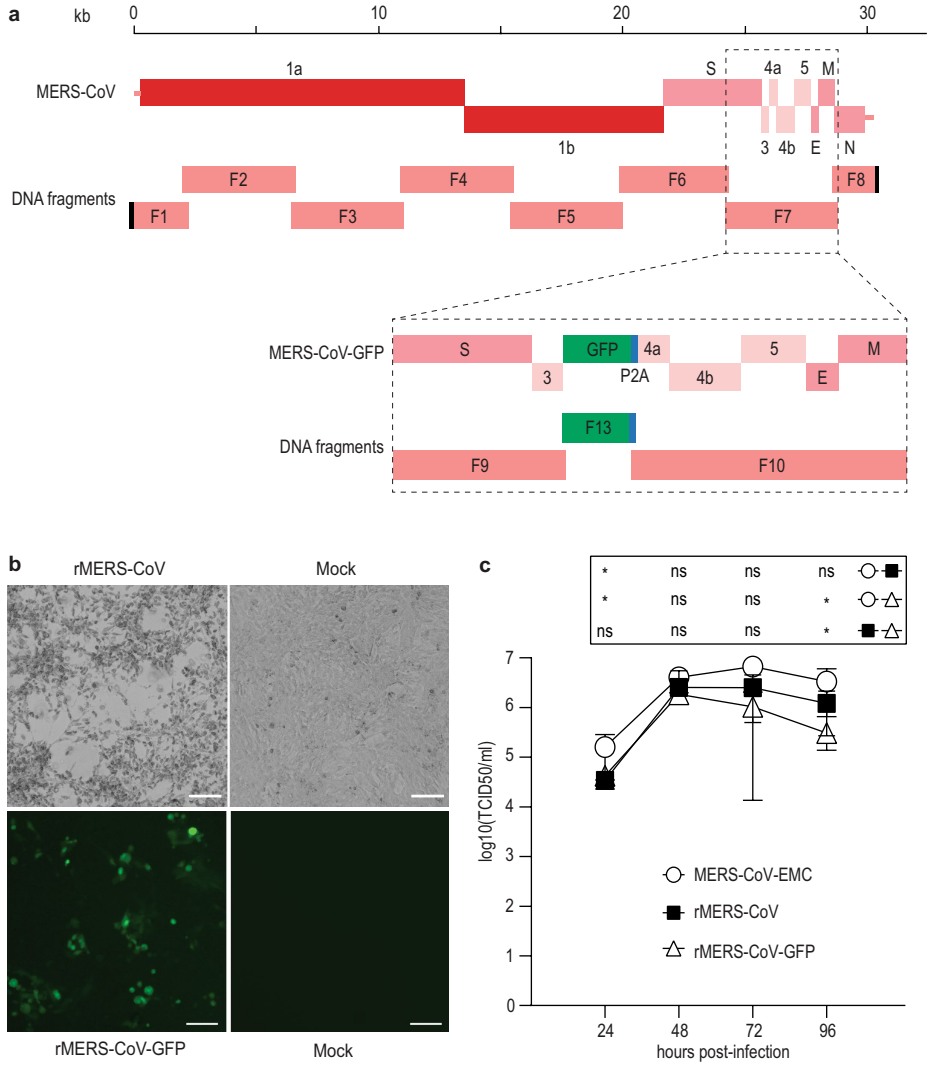
24. Raj, V. S. et al. Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. *Nature* **495**, 251–254 (2013).
25. Jonsdottir, H. R. et al. Establishment of primary transgenic human airway epithelial cell cultures to study respiratory virus–host interactions. *Viruses* **11**, 747 (2019).
26. Noskov, V. et al. A genetic system for direct selection of gene-positive clones during recombinational cloning in yeast. *Nucleic Acids Res.* **30**, E8 (2002).
27. Green, M. R. & Sambrook, J. *Molecular Cloning: A Laboratory Manual* 4th edn (Cold Spring Harbor Laboratory Press, 2012).
28. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).
29. Blount, B. A., Driessen, M. R. & Ellis, T. GC preps: fast and easy extraction of stable yeast genomic DNA. *Sci. Rep.* **6**, 26863 (2016).
30. Mutterer, J. & Zinck, E. Quick-and-clean article figures with FigureJ. *J. Microsc.* **252**, 89–91 (2013).
31. Hierholzer, J. C. & Killington, R. A. in *Virology Methods Manual* (eds Mahy, B. W. J. & Kangro, H. O.) 25–46 (Academic Press, 1996).
32. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
33. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
34. Woelfel, R. et al. Clinical presentation and virological assessment of hospitalized cases of coronavirus disease 2019 in a travel-associated transmission cluster. Preprint at medRxiv https://doi.org/10.1101/2020.03.05.20030502 (2020).
35. Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271–280 (2020).
36. Chan, J. F. et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microbes Infect.* **9**, 221–236 (2020).

Acknowledgements This work was supported by the European Commission (Marie Skłodowska-Curie Innovative Training Network "HONOURS"; grant agreement no. 721367), the Swiss National Science Foundation (SNF; grants CRSII3_160780 and 310030_173085), the German Research Council (DFG; grants SFB-TR84 (TRR 84/3, A07) and DR 772/7-2), the Federal Ministry of Education and Research (BMBF; grant RAPID, 01KI1723A) and by core funds of the University of Bern. We thank S. Vashee for the provision of the TAR vectors and for discussions related to his herpesvirus work; J. Peters Zocher for her advice in generating the figures; F. Suter-Riniker and P. Bittel for providing clinical samples; staff from the Next Generation Sequencing platform (University of Bern); and M. Schweizer, P. Plattet, M. Gerber, M. Friesland, M. Alves, N. Vielle, B. Zumkehr, M. Brügger and D. Brechbühl for reagents, technical advice and helpful discussions. This study is dedicated to S. Kunz.

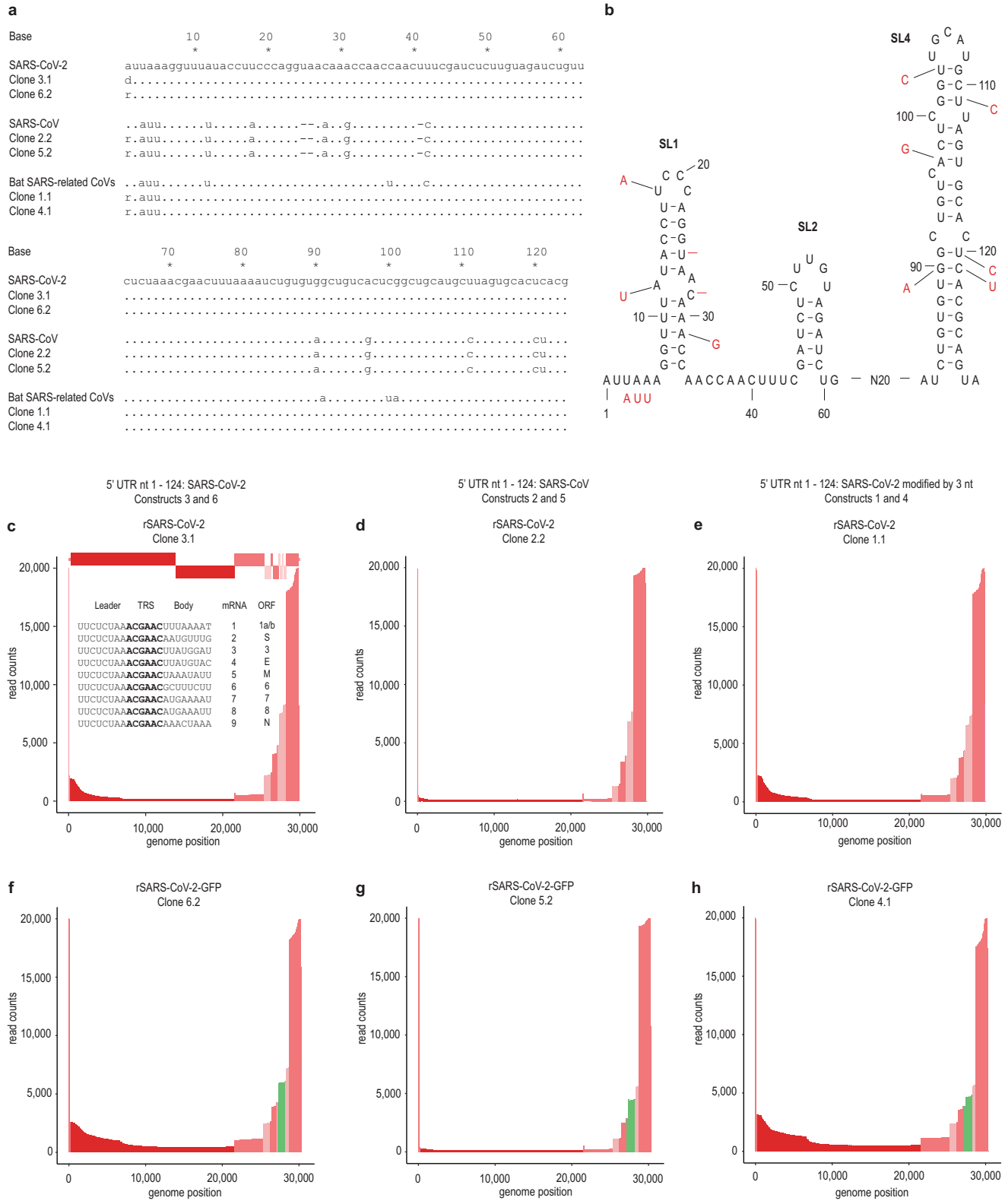
Author contributions V.T. and J.J. conceived the study. T.T.N.T., N.E. and F.L. performed most of the experiments. P.V., H.S., K.S., J.P., S. Steiner, M.H., A.K., M.G., L.L., L.H., M.W., S.P., D.H., V.C., S.C.-P., S. Schröder, D.M., D.N., V.M.C., M.A.M., C.D. and R.D. did experimental work and/or provided essential experimental systems and reagents. T.T.N.T., N.E., F.L., H.S., J.K. and R.D. performed sequencing analyses, including computational analyses. V.T., J.J., T.T.N.T., F.L. and N.E. wrote the manuscript and made the figures. All authors read and approved the final manuscript.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2294-9.
Correspondence and requests for materials should be addressed to J.J. or V.T.
Peer review information *Nature* thanks Luis Enjuanes, Stanley Perlman and Timothy Patrick Sheahan for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

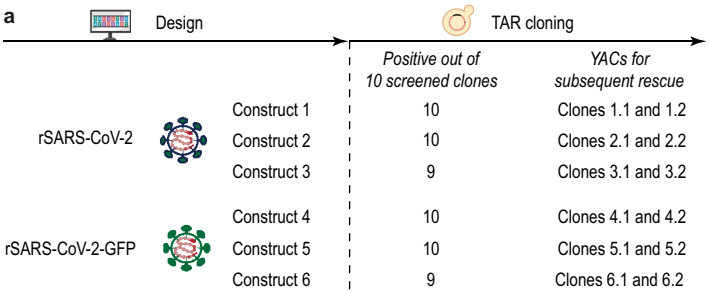


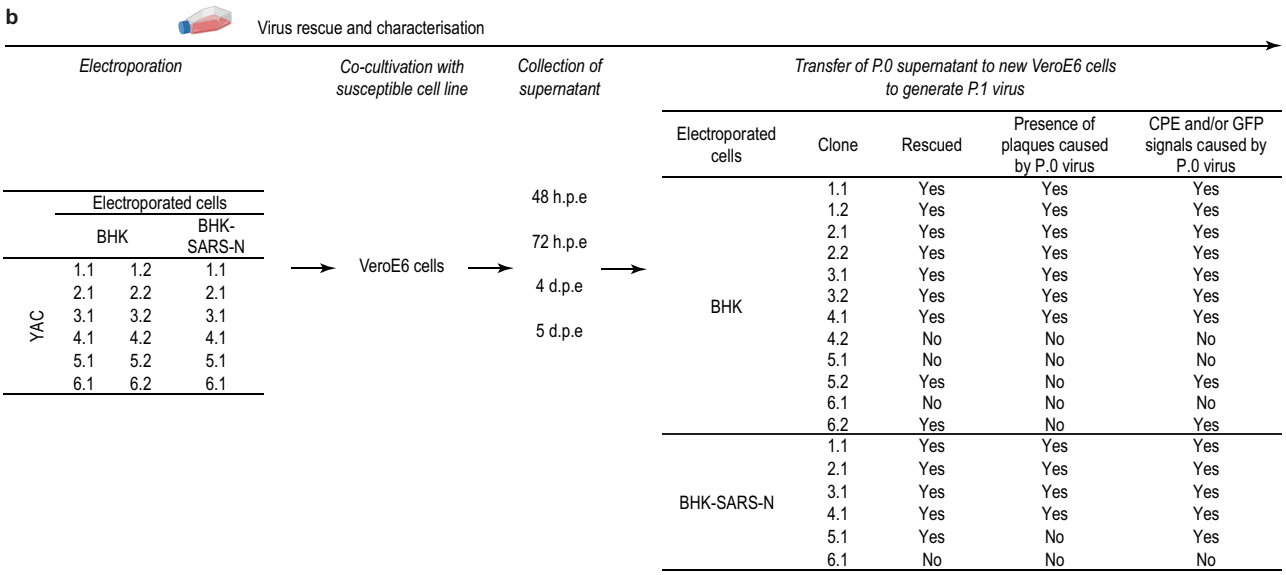
Extended Data Fig. 1 | Generation of viral cDNA clones and recovery of recombinant MERS-CoV and MERS-CoV-GFP. a, Schematic representation of the genome organization of MERS-CoV (top) and MERS-CoV-GFP (bottom) with 8 and 10 viral subgenomic overlapping fragments used for TAR cloning, respectively. **b**, Rescue of recombinant MERS-CoV and MERS-CoV-GFP. After the delivery of viral RNAs into BHK-21 cells using electroporation, the cells were co-cultured with Vero B4 cells, and supernatants containing recombinant viruses that were produced were used to infect new Vero B4 cells. Infected cells were visualized by bright-field microscopy for rMERS-CoV (top; 5 days after infection), and by fluorescence microscopy for GFP expression of rMERS-CoV-GFP (bottom, 3 days after infection). Mock, Vero B4 cells inoculated with the supernatant of BHK-21 cells that were electroporated without viral RNAs. Images are representative of two independent experiments. Scale bars, 100 µm. **c**, Replication kinetics of MERS-CoV-EMC, rMERS-CoV and rMERS-CoV-GFP. Vero B4 cells were infected (MOI = 0.01). Cell-culture supernatants were collected at the indicated time points after infection and titrated by $TCID_{50}$ assay. Data are the mean ± s.d. of three independent biological replicates. Statistical significance was determined by two-sided unpaired Student's *t*-test without adjustments for multiple comparisons. *P* values (from left to right): top, **P* = 0.0332; ns, *P* = 0.3294; ns, *P* = 0.2003; ns, *P* = 0.0966; middle, **P* = 0.0457; ns, *P* = 0.1233; ns, *P* = 0.0838; **P* = 0.0199; bottom, ns, *P* = 0.3240; ns, *P* = 0.6641; ns, *P* = 0.1376; **P* = 0.0427. TCID50/ml, 50% tissue culture infectious dose per ml.



Extended Data Fig. 2 | See next page for caption.

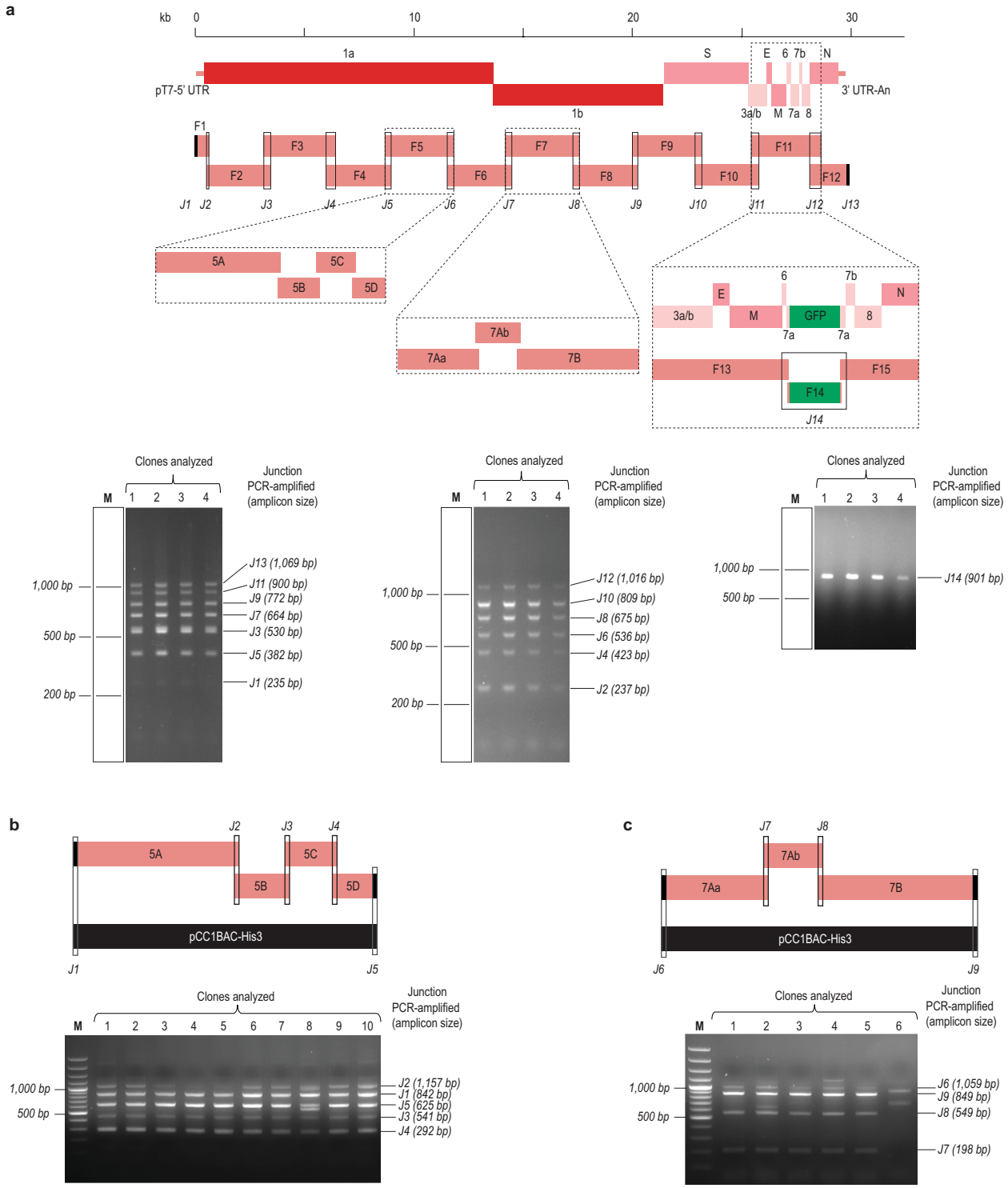
Extended Data Fig. 2 | Sequence analyses of the 5′ UTR of the SARS-CoV-2 genome. a, Sequence analysis using 5′RACE. Results from 5′RACE of rescued rSARS-CoV-2 and rSARS-CoV-2-GFP clones are shown as a sequence comparison of the first 124 nucleotides of the 5′UTR region of the SARS-CoV-2 genome (top; MN996528.1) used to generate clones 3.1 and 6.2; the SARS-CoV Frankfurt-1 isolate (middle; AY291315) used to generate clones 2.2 and 5.2 and bat SARS-related CoVs (bottom; ZXC21 and ZC45) used to generated clones 1.1 and 4.1. A 5′-RACE analysis has been performed from viral RNA for all clones and the sequence has been confirmed. **b**, Representation of predicted RNA stem-loop (SL) secondary structures within the 5′ UTR of SARS-CoV-2. The secondary structures of SARS-CoV-2 RNA were manually adjusted based on previously published RNA structure predictions[36]. Black letters and numbers represents the SARS-CoV-2 5′-terminal sequence. Red letters depict nucleotides that are different within the SARS-CoV 5′-terminal sequence (the '-' indicates a nucleotide deletion in SARS-CoV compared with SARS-CoV-2). N20 indicates 20 nucleotides. **c**–**h**, RNA-sequencing analysis of rSARS-CoV-2 clones 3.1 (**c**), 2.2 (**d**), 1.1 (**e**) and rSARS-CoV-2-GFP clones 6.2 (**f**), 5.2 (**g**), 4.1 (**h**). The sequence read coverage of the SARS-CoV-2 and SARS-CoV-2-GFP genomes is shown as read counts plotted according to the genome positions. The sequence read coverage is colour-coded according to the viral ORFs (red, ORF1a/b; dark pink, structural genes; light pink, accessory genes; green, GFP) to illustrate the characteristic pattern of the coronavirus transcription gradient of genomic and subgenomic viral RNAs. Leader–body junctions of viral RNAs were determined for SARS-CoV-2 clone 3.1 and are depicted in **c**. RNAs used for 5′-RACE (**a**) and RNA-sequencing analyses (**c**–**h**) were prepared from virus-infected Vero E6 cells (MOI = 0.001; 48 h after infection). The TRS is highlighted in bold.





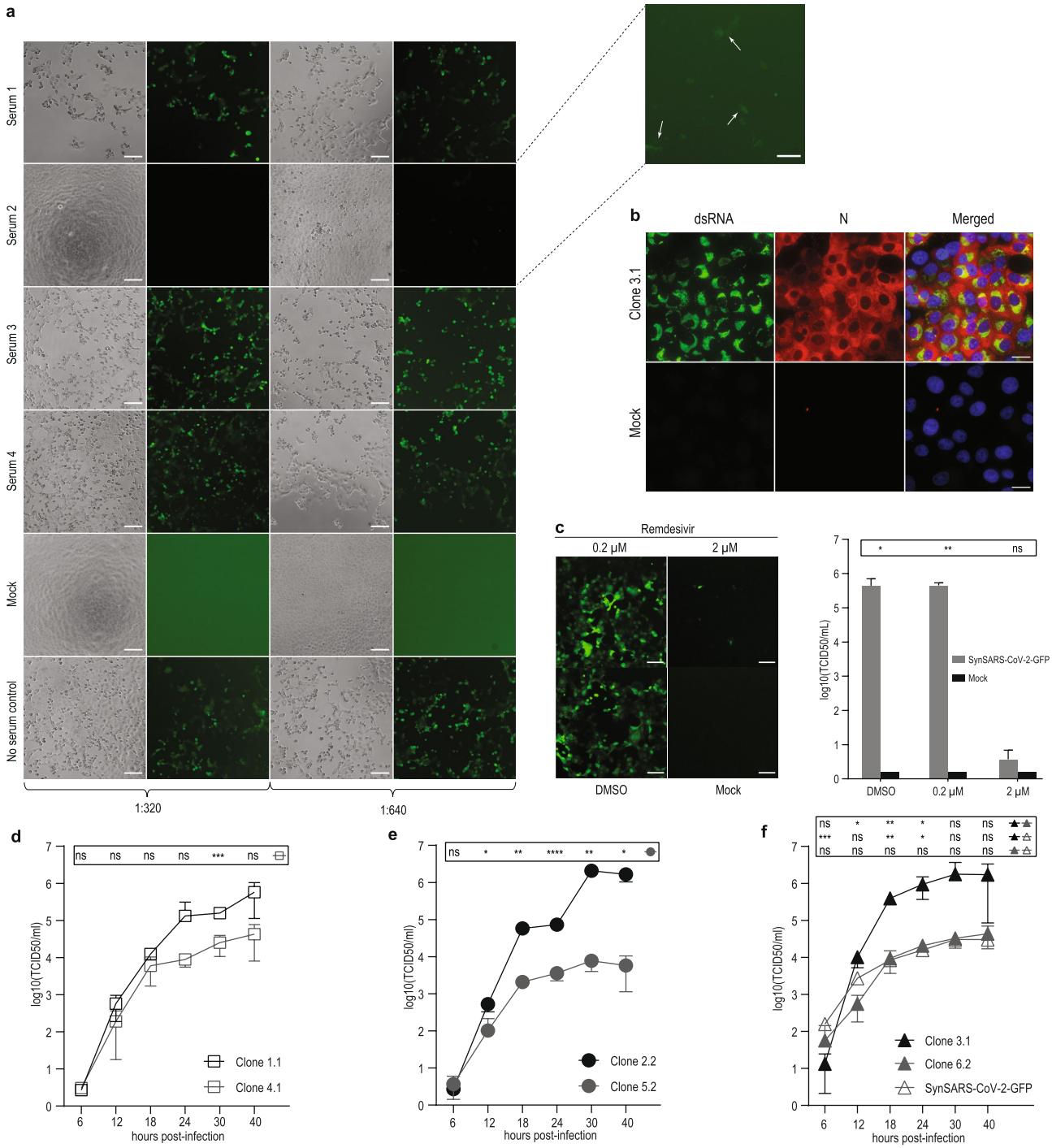
Electroporated cells	Clone	Rescued	Presence of plaques caused by P.0 virus	CPE and/or GFP signals caused by P.0 virus
BHK	1.1	Yes	Yes	Yes
	1.2	Yes	Yes	Yes
	2.1	Yes	Yes	Yes
	2.2	Yes	Yes	Yes
	3.1	Yes	Yes	Yes
	3.2	Yes	Yes	Yes
	4.1	Yes	Yes	Yes
	4.2	No	No	No
	5.1	No	No	No
	5.2	Yes	No	Yes
	6.1	No	No	No
	6.2	Yes	No	Yes
BHK-SARS-N	1.1	Yes	Yes	Yes
	2.1	Yes	Yes	Yes
	3.1	Yes	Yes	Yes
	4.1	Yes	Yes	Yes
	5.1	Yes	No	Yes
	6.1	No	No	No

Extended Data Fig. 3 | Workflow for the reconstruction and rescue of rSARS-CoV-2 and rSARS-CoV-2-GFP. a, Overview of the constructs and clones. Six constructs were initially designed on the basis of three different 5′-UTR regions. These regions comprised a modified sequence of the 5′-UTR region of SARS-CoV-2 (5′-AT**AUU**AGG) in which nucleotides 3–5 (UAA) of SARS-CoV-2 were changed to AUU to match nucleotides 3–5 of bat SARS-related CoV (constructs 1 and 4); a SARS-CoV-2 5′-terminus in which the first 124 nucleotides were changed to the corresponding 5′-terminal sequence of SARS-CoV (constructs 2 and 5); and the reported sequence of the SARS-CoV-2 (MN996528.1) (constructs 3 and 6). After transformation in yeast, ten colonies were randomly picked for each of the six constructs and all of the junctions bridging the overlapping fragments were verified by multiplex PCR. For each construct, two clones (*x*.1 and *x*.2) were randomly selected and YAC DNAs were isolated (12 clones in total). **b**, Rescue of rSARS-CoV-2 and rSARS-CoV-2-GFP clones. RNAs were generated from YAC DNAs by in vitro transcription and electroporated together with an mRNA encoding the SARS-CoV-2 N protein either into BHK-21 cells (12 clones) or BHK-SARS-N cells (cells expressing the SARS-CoV N protein) (6 clones). Electroporated cells were then co-cultured with susceptible Vero E6 cells to rescue the recombinant viruses. Passage 0 (P.0) supernatants were collected at different time points after electroporation (from 2 to 5 days after electroporation) and transferred to Vero E6 cells to generate passage 1 (P.1) virus stocks, and in parallel to demonstrate the presence of infectious virus in plaque assays (for virus clones that do not encode GFP) or fluorescence microscopy (for GFP-encoding virus clones). h.p.e, hours post-electroporation; d.p.e, days post-electroporation; CPE, cytopathogenic effects.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Reconstruction of synSARS-CoV-2-GFP and TAR cloning of full-length synthetic fragments 5 and 7 in yeast. a, Genome organization of the synSARS-CoV-2-GFP and 19 viral fragments used for TAR cloning (F1–F10; F12–F15). Fragments 5 and 7 were split in four (5A–5D) and three (7Aa, 7Ab, 7B) DNA parts, respectively. Viral ORFs, the ORF for GFP and sequence elements at the 5′ UTR and 3′ UTR are indicated. Primers used to generate the fragments are listed in Supplementary Table 1. J2–J12 and J14 represent the junctions, that is, overlapping regions, between the subgenomic fragments. J1 and J13 represent junctions with the TAR vector. Gel images show the results of two multiplex PCRs designed to confirm the presence of correctly recombined junctions. Multiplex PCR using set 1 primers (left) detects junctions J1, J3, J5, J7, J9, J11 and J13, and multiplex PCR using set 2 primers (middle) detects junctions J2, J4, J6, J8, J10 and J12. The presence of the *GFP* gene inserted in fragment 14 was confirmed (right). PCR-product sizes are depicted and confirm the proper assembly of the synSARS-CoV-2 full-length genome in all four YAC clones analysed. **b**, TAR cloning of the full-length synthetic fragment 5 in yeast. Four overlapping synthetic DNA fragments (5A–5D) provided by Genscript were reconstructed as a YAC after transformation in yeast. Correct reassembly was confirmed by multiplex PCR over junctions J1–J5 for 9 out of the10 clones screened (clone 8 was considered incorrect). **c**, TAR cloning of the full-length synthetic fragment 7 in yeast. Full-length fragment 7 was assembled by TAR cloning using 3 synthetic dsDNA parts (7Aa, 7Ab and 7B) provided by Genscript. Correct reassembly was confirmed by multiplex PCR over junctions J6–J9 for 5 out of 6 clones (clone 6 is considered incorrect). Cloning experiments shown in **a**–**c** have been performed once. pT7, T7 RNA polymerase promoter; An, poly(A) tail; M, GeneRuler 100-bp plus DNA marker (Thermo Scientific).



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Characterization of rSARS-CoV-2, rSARS-CoV-2-GFP and synSARS-CoV-2-GFP. a, Serum neutralization assay showing the use of synSARS-CoV-2-GFP. Inactivated sera were serially diluted and incubated with synSARS-CoV-2-GFP (250 $TCID_{50}$) for 1 h before infection of Vero E6 cells. Results at 48 h after infection show that at the virus dilution of 1:320 in serum, virus infection was not observed in serum 2 (convalescent human anti-SARS-CoV-2 serum). At the following dilution (1:640), GFP expression was detected, but cytopathogenic effects were not yet detected. Sera 1 and 3, control sera; serum 4, convalescent human anti-SARS-CoV serum[35]. The serum neutralization experiment was performed twice with similar results. Mock, uninfected cells; no serum control, infected cells with no serum added. **b**, Immunofluorescence assay showing viral protein synthesis. Vero E6 cells were infected with rSARS-CoV-2 clone 3.1 (passage 2) (MOI = 0.01). Mock, uninfected cells. At 48 h after infection, cells were fixed and prepared for immunofluorescence staining with primary antibodies directed against dsRNA and SARS-CoV nucleocapsid (N). Green, dsRNA; red, viral N protein; blue, DAPI. The experiment was performed twice with different Vero E6 cells (passages 10 and 11) with similar results. **c**, Remdesivir treatment showing the use of synSARS-CoV-2-GFP. Vero E6 cells were infected with synSARS-CoV-2-GFP (MOI = 0.01) and treated with 0.2 µM, 2 µM or without remdesivir. DMSO was used as treatment in cells that were not incubated with remdesivir. Mock, uninfected cells. At 48 h after infection, cells were analysed by fluorescence microscopy to detect GFP expression (left) and cell-culture supernatants were collected and titrated by $TCID_{50}$ assay (right). Titration data represent the mean ± s.d. of three independent biological replicates ($n = 3$). Statistical significance was determined for synSARS-CoV-2 compared with mock by two-sided unpaired Student's *t*-test without adjustment for multiple comparisons. *P* values (from left to right): $^{*}P = 0.0460$; $^{**}P = 0.0010$; ns, $P = 0.2972$. Scale bars, 100 µm (**a**–**c**). **d**–**f**, Comparison of replication kinetics of rSARS-CoV-2, rSARS-CoV-2-GFP and synSARS-CoV-2-GFP. Data are from the same experiment as shown in Fig. 3c but each panel now shows the direct comparison of rSARS-CoV-2, rSARS-CoV-2-GFP and synSARS-CoV-2-GFP clones that have the same 5′ terminus: clones 1.1 and 4.1 (5′-AU**AUU**AGG) (**d**), clones 2.2 and 5.2 (124 5′-terminal nucleotides of SARS-CoV) (**e**) and clones 3.1, 6.2 and synSARS-CoV-2-GFP (authentic SARS-CoV-2 sequence (5′-AU**UAA**AGG) according to MN996528.1) (**f**). Vero E6 cells were infected (MOI = 0.01) and supernatants were collected at indicated time points after infection and titrated by $TCID_{50}$ assay. Data represent the mean ± s.d. of three independent biological replicates. Statistical significance was determined by two-sided unpaired Student's *t*-test without adjustments for multiple comparisons. **d**, *P* values (from left to right): ns, $P = 0.5182$; ns, $P = 0.1920$; ns, $P = 0.0993$; ns, $P = 0.3001$; $^{***}P = 9.1 \times 10^{-5}$. **e**, *P* values (from left to right): ns, $P = 0.4978$; $^{*}P = 0.0324$; $^{**}P = 0.0020$; $^{****}P < 10^{-6}$; $^{**}P = 0.0011$; $^{*}P = 0.1000$. **f**, *P* values (from left to right): top, ns, $P = 0.4427$; $^{*}P = 0.02474$; $^{**}P = 0.0019$; $^{*}P = 0.0490$; ns, $P = 0.1867$; ns, $P = 0.1502$; middle, $^{***}P = 2.4 \times 10^{-5}$; ns, $P = 0.1109$; $^{**}P = 0.0019$; $^{*}P = 0.4825$; ns, $P = 0.1862$; ns, $P = 0.1478$; last, ns, $P = 0.1161$; ns, $P = 0.4026$; ns, $P = 0.8700$; ns, $P = 0.1161$; ns, $P = 0.8626$; ns, $P = 0.4502$.

Extended Data Table 1 | List of restriction enzymes used to linearize the YAC DNA for each virus before in vitro transcription

Virus	Restriction enzyme
MHV-GFP	*Pac*I
MERS-CoV	*Mlu*I
MERS-CoV-GFP	*Mlu*I
HCoV-229E	*Eag*I
HCoV-HKU1	*Eag*I
MERS-CoV-Riyadh-1734-2015	*Mlu*I
ZIKA virus	*Pac*I
Human RSV-B	*n.a*
SARS-CoV-2	*Eag*I
SARS-CoV-2-GFP	*Eag*I

n.a., not applicable.

Extended Data Table 2 | Stability of the MHV-GFP and MERS-CoV YAC clones in yeast

	YAC clones							
	MHV-GFP				MERS-CoV			
	Clone 1		Clone 2		Clone 1		Clone 2	
	Passage 1	Passage 17	Passage 1	Passage 17	Passage 1	Passage 15	Passage 1	Passage 15
Non-native viral sequence(s)	none	none	none	none	none	none	none	none
Mutation(s) due to serial passaging	none	none	none	none	none	none	none	none
Mutation(s) due to RT-PCR	1	1	none	none	none	none	none	none

Extended Data Table 3 | Details of the synthetic HCoV-HKU1, SARS-CoV-2, MERS-CoV Riyadh-1734-2015 and ZIKA virus fragments

Fragment	Genome position (5' UTR to 3' UTR)		Fragment size (bp)	Cloning vector	Antibiotic selection	Restriction sites		
	5'-end	3'-end				5'-end	3'-end	Vector digestion
HCoV-HKU1								
Fragment 1	1	2399	2506	pUC57	Amp^R			
Fragment 2	8526	10920	2395	pUC57	Amp^R			
Fragment 3	19107	21530	2424	pUC57	Amp^R			
Fragment 4	27484	29925	2646	pUC57	Amp^R			
SARS-CoV-2								
Fragment 1.1	1	483	579	pUC57	Amp^R	*Sma*I	*Sma*I	
Fragment 1.2	1	483	579	pUC57	Amp^R	*Sma*I	*Sma*I	
Fragment 1.3	1	483	579	pUC57	Amp^R	*Sma*I	*Sma*I	
Fragment 2	377	3325	2949	pUC57	Amp^R	*Sma*I	*Sma*I	*Pvu*I
Fragment 3	3012	6315	3304	pUC57	Amp^R	*Eco*RV	*Eco*RV	*Pvu*I
Fragment 4	6003	8994	2992	pUC57mini	Amp^R	*Eco*RV	*Nde*I	
Fragment 5	8718	11966	3249	pCC1-His3	$Chlo^R$	*Sma*I	*Sma*I	
Fragment 5A	8718	10475	1757	pUC19	Amp^R	*Sma*I		
Fragment 5B	10446	11023	573	pUC19	Amp^R			
Fragment 5C	10994	11540	573	pUC19	Amp^R			
Fragment 5D	11511	11966	456	pUC19	Amp^R		*Sma*I	
Fragment 6	11664	14605	2942	pUC57mini	Amp^R	*Sma*I	*Sma*I	
Fragment 7	14311	17698	3388	pCC1-His3	$Chlo^R$	*Sma*I	*Sma*I	
Fragment 7Aa	14311	15445	1134			*Sma*I		
Fragment 7Ab	15416	16029	614					
Fragment 7B	16000	17698	1699	pUC57	Amp^R		*Sma*I	
Fragment 8	17399	20358	2960	pUC57mini	Amp^R	*Sma*I	*Sma*I	
Fragment 9	20110	23286	3177	pUC57mini	Amp^R	*Sma*I	*Sma*I	
Fragment 10	22975	25940	2966	pUC57mini	Amp^R	*Sma*I	*Sma*I	
Fragment 11	25595	28779	3185	pUC57mini	Amp^R	*Sma*I	*Sma*I	
Fragment 12	28274	29870	1812	pUC57	Amp^R	*Sma*I	*Sma*I	
pCC1-MERS-CoV-Riyadh-1734-2015								
Fragment 1	1	2190	2339					
Fragment 8	28447	30126	1791	pUC57	Amp^R			
ZIKA virus								
Fragment 6	10627	10807	298	pUC57	Amp^R			

Extended Data Table 4 | Mutations in rSARS-CoV-2(-GFP) YAC clones after sequence confirmation by Sanger sequencing

	Mutation location		Mutation					Clones													
No	Genome position	Synthetic fragment	REF	ALT	Type	Codon change	Subclones: AA change	1.1	1.2	2.1	2.2	3.1	3.2	4.1	4.2	5.1	5.2	6.1	6.2	Syn.	Munich isolate
	8987		T	C	SNP	TTT->CTT	Phe->Leu	T	C	T	T	T	T	T	T	T	T	T	T	T	T
2	9143		T	C	SNP	TCT->CCT	Ser->Pro	T	T	T	T	T	T	T	T	T	T	C	T	T	T
3	10845	Fragment 5	T	C	SNP	ATG->ACG	Met->Thr	T	T	T	T	T	T	T	C	T	T	T	T	T	T
4	11337		T	C	SNP	TTA->TCA	Leu->Ser	T	T	T	T	T	T	T	T	T	T	C	T	T	T
5	11571		T	C	SNP	TTC->TCC	Phe->Ser	T	T	T	T	T	T	T	C	T	T	T	T	T	T
6	11811		G	A	SNP	GGC->GAC	Gly->Asp	G	G	G	G	G	G	G	A	G	G	G	G	G	G
7	14580		G	A	SNP	ATG->ATA	Met->Ile	G	G	G	G	G	G	G	G	G	G	A	G	G	G
8	14767		G	A	SNP	GCT->ACT	Ala->Thr	G	G	G	G	G	G	G	G	A	G	G	G	G	G
9	15531		T	C	SNP	TGT->TGC	none (Cys)	T	T	T	T	T	T	T	T	T	T	C	T	T	T
10	15819	Fragment 7	T	Ø	Del	TTA->TAT	Leu->Tyr (frameshift)	T	T	T	T	T	T	T	T	T	T	Ø	T	T	T
11	17197		G	A	SNP	GCA->ACA	Ala->Thr	G	A	G	G	G	G	G	G	G	G	G	G	G	G
12	17432		T	G	SNP	ATT->AGT	Lys->Ser	T	T	T	T	T	G	T	T	T	T	T	T	T	T
13	17539		G	A	SNP	GAC->AAC	Asp->Asn	G	G	G	G	G	G	G	G	G	G	G	A	G	G

Genome position indicates the position in the rSARS-CoV-2 genome (MN996528.1). REF indicates the sequence in the rSARS-CoV-2 genome; ALT indicates variation in the rSARS-CoV-2 YAC clone. Codon change shows the effect of the mutation on the corresponding codon. AA change lists the effect of the mutation on the corresponding amino acid. Del, deletion; SNP, single-nucleotide polymorphism; Syn, synSARS-CoV-2-GFP.

nature research

Corresponding author(s): Volker Thiel

Last updated by author(s): 2020/04/15

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data acquisition for Minion sequencing was done with MinKNOW version 19.06.9 and basecalling software Guppy version 3.4.5.
Data analysis	seqeunce analysis: Geneious Prime ® 2019.2.3 virus kinetics analyses: GraphPad Prism version 8.3.0 for Windows figures: Adobe Illustrator and Biorender IFA: FIJI with FigureJ plugin NGS (RNAseq): TrimGalore software (version 0.6.5), STAR (version 2.7.0a); SAMtools (version 1.10); Minion sequencing: Python command-line qcat (Mozilla Public License 2.0. Copyright © 2018 Oxford Nanopore Technologies Ltd. qcat (v1.1.0); Minimap2 (Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All Data and files will be made available. The following genome sequences have been submitted to GenBank: rSARS-CoV-2 (#MT108784), hRSV/B/Bern/2019 (#MT107528); MERS-CoV-Riyadh-1734-2015 (#MN481979). The RNAseq data of rSARS-CoV-2(-GFP) has been submitted to the NCBI Sequence Read Archive

(BioProject ID number: PRJNA615319; BioSample accessions: SAMN14450686, SAMN14450687, SAMN14450688, SAMN14450689, SAMN14450690, SAMN14450691).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed. Sample sizes were based on standards in the field, typically 3 independent biological replicates, with each replicate assayed in technical duplicate or triplicate.
Data exclusions	no data was excluded
Replication	all attempts at replication were successful; experiments were performed according to best practices and as described in the methods.
Randomization	randomization was not applied since cloning procedures, virus infection/titrations, and inhibitor/neutralization experiments did not require randomization.
Blinding	blinding was done for remdsivir inhibition assay and virus neutralisation assay to ensure that images taken from infected cultures are representative.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	anti-SARS-CoV Nucleocapsid (N) protein (rabbit), Rockland, Product No: 200-401-50, Lot No 16570, dillution 1:1000; anti-dsRNA, J2, English and Scientific Consulting, Product No:10010500, clone J2, Lot No J2-1913, dillution 1:200.
Validation	anti-SARS-CoV Nucleocapsid (N) protein (rabbit) and anti-dsRNA, J2 were validated by comparing infected vs uninfected cells and by assessing various dillutions.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	L929 cells (Source : 85011425 (Sigma, ECACC, 2017); authenticated 04/2019); 17Cl-1 cells (gift from Stanley Sawicki; authenticated 04/2019); Huh7 cells (gift from Volker Lohman, University of Heidelberg; authetication was done in Heidelberg) Vero, VeroB4 and VeroE6 (obtained from Marcel Müller, Charité, Berlin (co-author); authentication done in Berlin
Authentication	Profiling of cell line was done using highly-polymorphic short tandem repeat loci (STRs). Fragment analysis was done on an ABI3730xl (Life Technologies) and the resulting data were analyzed with GeneMarker software (Softgenetics).

Mycoplasma contamination	all cell lines in our laboratory are routinely screened for micoplasma contamination and were tested negative.
Commonly misidentified lines (See ICLAC register)	none

Article

A metabolic pathway for bile acid dehydroxylation by the gut microbiome

https://doi.org/10.1038/s41586-020-2396-4

Received: 27 August 2019

Accepted: 18 March 2020

Published online: 17 June 2020

Check for updates

Masanori Funabashi[1,6,7], Tyler L. Grove[2,7], Min Wang[1], Yug Varma[1], Molly E. McFadden[3], Laura C. Brown[3], Chunjun Guo[1], Steven Higginbottom[4], Steven C. Almo[2]✉ & Michael A. Fischbach[1,5]✉

The gut microbiota synthesize hundreds of molecules, many of which influence host physiology. Among the most abundant metabolites are the secondary bile acids deoxycholic acid (DCA) and lithocholic acid (LCA), which accumulate at concentrations of around 500 μM and are known to block the growth of *Clostridium difficile*[1], promote hepatocellular carcinoma[2] and modulate host metabolism via the G-protein-coupled receptor TGR5 (ref. [3]). More broadly, DCA, LCA and their derivatives are major components of the recirculating pool of bile acids[4]; the size and composition of this pool are a target of therapies for primary biliary cholangitis and nonalcoholic steatohepatitis. Nonetheless, despite the clear impact of DCA and LCA on host physiology, an incomplete knowledge of their biosynthetic genes and a lack of genetic tools to enable modification of their native microbial producers limit our ability to modulate secondary bile acid levels in the host. Here we complete the pathway to DCA and LCA by assigning and characterizing enzymes for each of the steps in its reductive arm, revealing a strategy in which the A–B rings of the steroid core are transiently converted into an electron acceptor for two reductive steps carried out by Fe–S flavoenzymes. Using anaerobic in vitro reconstitution, we establish that a set of six enzymes is necessary and sufficient for the eight-step conversion of cholic acid to DCA. We then engineer the pathway into *Clostridium sporogenes*, conferring production of DCA and LCA on a nonproducing commensal and demonstrating that a microbiome-derived pathway can be expressed and controlled heterologously. These data establish a complete pathway to two central components of the bile acid pool.

The human gut microbiota harbour hundreds of metabolic pathways, most of which are encoded by genes that have not yet been identified[5–8]. Their small-molecule products are of interest for three reasons. First, most derive predominantly or exclusively from the microbiota (that is, there is no host source), and many enter the circulation, where they can have effects on peripheral tissues and organ systems. Second, their concentrations are similar to or exceed those of a typical drug; for example, indoxyl sulfate can accumulate in the human host at 130 mg per day[9]. Moreover, their concentration ranges are large, typically more than tenfold[10], which could help to explain microbiome-mediated biological differences among people. Finally, of the few high-abundance molecules whose biological functions are well understood, most are ligands for a key host receptor; for example, short-chain fatty acids modulate host immune function via GPR41/GPR43 (refs. [11–13]). Thus, high-abundance, microbiota-derived molecules are responsible for a remarkably broad range of phenotypes conferred on the host by bacteria.

Among these pathways, 7α-dehydroxylation of the primary bile acids cholic acid and chenodeoxycholic acid (CDCA) is particularly notable because the organisms that carry it out are present at very low abundance—an estimated ratio of 1:10[6] in a typical gut community[14]—yet they fully process a pool of primary bile acids that reaches concentrations of about 1 mM (ref. [15]). Therefore, the flux through this pathway must be very high in the small subset of cells in which it operates, and the low-abundance organisms in the microbiome that perform this transformation have an unusually large impact on the pool of metabolites that enters the host. This pathway's products—DCA and LCA—are the most abundant secondary bile acids in humans (up to 450–700 μM in caecal contents)[16], and are known to be important in three biological contexts: prevention of *C. difficile* outgrowth[1], induction of hepatocellular carcinogenesis[2], and modulation of host metabolic and immune responses[17–19]. More broadly, DCA, LCA and their derivatives are a major component of the recirculating bile acid pool, representing more than 90% of the pool in the intestine and more than 25% in the gallbladder[15]. These microbiome-derived bile acids are therefore central to understanding the efficacy of therapeutics that target the bile acid pool and are approved or in clinical

[1]Department of Bioengineering and ChEM-H, Stanford University, Stanford, CA, USA. [2]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY, USA. [3]Department of Chemistry, Indiana University, Bloomington, IN, USA. [4]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA. [5]Chan Zuckerberg Biohub, San Francisco, CA, USA. [6]Present address: Translational Research Department, Daiichi Sankyo RD Novare Co. Ltd, Tokyo, Japan. [7]These authors contributed equally: Masanori Funabashi, Tyler L. Grove. ✉e-mail: steve.almo@einstein.yu.edu; fischbach@fischbachgroup.org

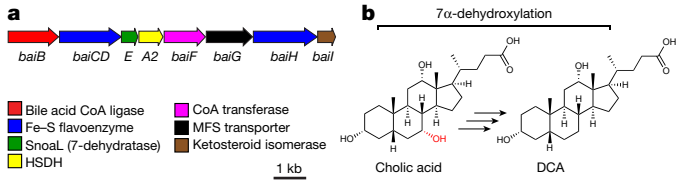


Fig. 1 | Schematics showing the *bai* operon and 7α-dehydroxylation. **a**, The *bai* operon consists of eight genes: seven encode enzymes and the eighth, *baiG*, encodes a transporter. It is conserved in every bacterial species known to 7α-dehydroxylate primary bile acids, and its gene products have been linked to specific steps in the pathway. HSDH, hydroxysteroid dehydrogenase. **b**, A simplified schematic showing the dehydroxylation of cholic acid to DCA.

trials for treatment of primary biliary cholangitis and nonalcoholic steatohepatitis[4].

In 1980 it was shown that the gut bacterium *Clostridium scindens* VPI 12708 carries out the 7α-dehydroxylation of cholic acid to produce DCA[20]. The knowledge that cholic acid serves as an inducer of 7α-dehydroxylation led to the discovery of a bile-acid-induced operon (termed *bai*) containing eight genes (Fig. 1 and Extended Data Fig. 1)[21]. Through heterologous expression and characterization of individual *bai* gene products, enzymes have been attributed to each step of the oxidative arm of the pathway[22–27], but the reductive arm remained poorly characterized[28]. A complete understanding of the pathway would enable efforts to control the composition of the bile acid pool by engineering the microbiome.

Here, by purifying and assaying pathway enzymes under anaerobic conditions, we reconstituted 7α-dehydroxylation in vitro. We then transferred the pathway from its genetically intractable producer *C. scindens* into *C. sporogenes*, conferring production of DCA and LCA on a nonproducing commensal bacterial species. These data establish a complete pathway for two central components of the bile acid pool, and they provide a genetic basis for controlling the bile acid output of the microbiome.

Reconstitution of 7α-dehydroxylation

We first set out to de-orphan the remaining steps in the 7α-dehydroxylation pathway. Because previous studies of the *bai* enzymes involved expressing them individually in *Escherichia coli*, we reasoned that an alternative approach—in which enzymes are purified, mixed and assayed in vitro—could help to delineate the set of enzymes necessary and sufficient for 7α-dehydroxylation. Given that the eight-gene *bai* operon is shared among all known 7α-dehydroxylating strains, we focused our efforts on the enzymes encoded by the operon. We cloned three orthologues of each enzyme, expressed them individually in *E. coli* under microaerobic conditions, and purified them anaerobically as amino-terminal His_6 fusions. Using this strategy, we obtained at least one soluble, purified orthologue of each Bai enzyme (Extended Data Fig. 2). When we incubated a mixture of the purified Bai enzymes with cholic acid, nicotinamide adenine dinucleotide $(NAD)^+$, coenzyme A and ATP under anaerobic conditions and monitored the reaction by liquid chromatography with mass spectrometry (LC–MS), we observed the time-dependent conversion of cholic acid to DCA, indicating that the combination of BaiB, BaiCD, BaiA2, BaiE, BaiF, and BaiH is sufficient for 7α-dehydroxylation; no additional enzymes are required (Fig. 2a, b).

To test our hypotheses regarding the order of steps in the pathway, we performed stepwise reconstitutions in which enzymes were added one at a time and intermediates were allowed to build up at each step in the pathway (Fig. 2c). From these data, we draw two conclusions. First, the six enzymes used in the reconstitution are not just sufficient but also necessary, and the pathway proceeds according to the scheme shown in Fig. 2c. We directly observed mass ions consistent with each of the proposed intermediates, providing direct evidence for the previously proposed portion of the biosynthetic route. (See Supplementary Table 1 and Extended Data Fig. 3 for data supporting our provisional structural assignments; two important limitations are that we do not have authentic standards for all intermediates, and that the ability to distinguish bile acid isomers by LC–MS can be limited.) In spite of its conservation in all known dehydroxylating species, BaiI is dispensable for cholic acid dehydroxylation in vitro. As BaiI is a predicted Δ^5-ketosteroid isomerase, it may process a substrate other than cholic acid, probably one with a 4,5- or 5,6-olefin.

Second, to our surprise, the absence of BaiH caused the pathway to stall at the highly oxidized intermediate 3-oxo-4,5-6,7-didehydro-DCA, and its addition resulted in two successive $2e^-$ reductions to form 3-oxo-DCA. BaiH had previously been proposed to oxidize an alternative substrate, 3-oxo-4,5-dehydro-ursodeoxycholic acid[25], so a potential role in the reductive arm of the pathway was unexpected. To explore this finding further, we incubated purified BaiH with synthetic 3-oxo-4,5-6,7-didehydro-DCA; we observed that the enzyme catalyses a $2e^-$ reduction to 3-oxo-4,5-dehydro-DCA, but does not reduce this intermediate further (Extended Data Fig. 4). Notably, 3-oxo-4,5-dehydro-DCA does not build up in the reconstitution reaction containing BaiH, suggesting that another enzyme present in the mixture catalyses the second reductive step. Hypothesizing that the BaiH homologue BaiCD catalyses the second reductive step, we incubated it with synthetic 3-oxo-4,5-dehydro-DCA, revealing that it reduces this substrate to 3-oxo-DCA (Extended Data Fig. 4). Together, these data show that the pathway uses an unusual redox strategy in which the A and B rings of the steroid core are converted into a highly oxidized intermediate, 3-oxo-4,5-6,7-didehydro-DCA; and that the two key reductive steps are catalysed by two homologous enzymes in the Fe–S flavoenzyme superfamily, BaiH and BaiCD.

Finally, the last step in the pathway—reduction of 3-oxo-DCA to DCA—is carried out by BaiA2, as confirmed by assaying purified BaiA2 alone (Extended Data Fig. 5). Thus, BaiA2 and BaiCD both act twice in the pathway, catalysing its first two and last two redox steps.

Engineering the pathway into *C. sporogenes*

Having determined the set of enzymes that are necessary and sufficient for the pathway, we sought to gain genetic control over the pathway as a first step towards engineering the bile acid output of the gut community. We began by attempting to construct a mutation in the *baiCD* gene of the native producer, *C. scindens*, using the ClosTron group II intron system; however, we were unsuccessful owing to an inability to introduce DNA constructs into *C. scindens* by conjugation. As an alternative approach, we considered expressing the *bai* pathway in a gut commensal that is unable to carry out 7α-dehydroxylation; however, with notable exceptions[29–32], methods for transferring pathways in *Clostridium* are underdeveloped. To our knowledge, no pathway from the human microbiome has been mobilized from one *Clostridium* species to another.

We selected *C. sporogenes* American Type Culture Collection (ATCC) strain 15579 as the recipient for two reasons: it is related to *C. scindens*, making it likely that ancillary metabolic requirements for the pathway (for example, cofactor biogenesis) would be met; and genetic tools have been developed that enable plasmids to be transformed into *C. sporogenes*[33]. Our initial attempts to clone the entire eight-gene *bai* operon (*baiB–baiI*) into an *E. coli–C. sporogenes* shuttle vector failed to yield clones harbouring the complete operon. Reasoning that there might be a gene in the cluster that is toxic to *E. coli*, we cloned various fragments of the cluster under the control of different promoters (detailed in Supplementary Table 2), eventually managing to split the cluster into three pieces, each in its own *E. coli–C. sporogenes* shuttle vector: *baiB–baiF* in pMTL83153 (pMF01), *baiG* in pMTL83353 (pMF02),

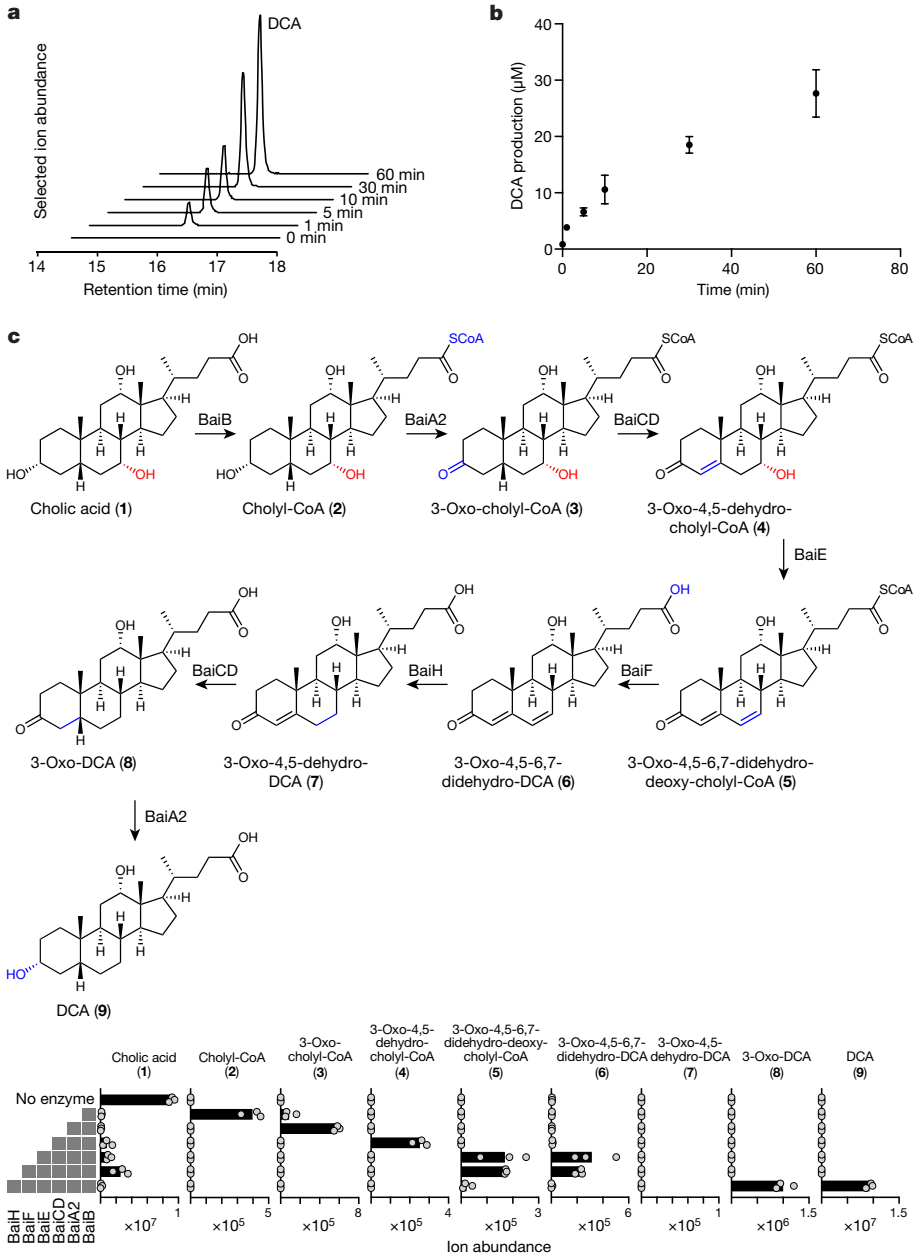


Fig. 2 | Establishing the complete 7α-dehydroxylation pathway in vitro. **a**, Extracted ion chromatograms (EICs) showing time-dependent production of DCA by six purified Bai enzymes. BaiB, BaiCD, BaiA2, BaiE, BaiF and BaiH were purified and assayed anaerobically in the presence of NAD^+, CoA, and ATP. Reactions were initiated by adding cholic acid, and aliquots were analysed by LC–MS at the indicated time points. The experiment was repeated twice independently with similar results. **b**, Time course of DCA production by a mixture of BaiB, BaiCD, BaiA2, BaiE, BaiF and BaiH. Data points indicate the mean level of DCA ± s.d. (three biological replicates). **c**, Top, complete proposed pathway for the 7α-dehydroxylation of cholic acid to DCA. Bottom, LC–MS ion abundance for DCA (**9**) and pathway intermediates (**1**–**8**) produced by a stepwise reconstitution assay in which the indicated enzymes were co-incubated as described in **a**. Bars show means of three independent biological replicates.

and *baiH–baiI* in pMTL83253 (pMF03) (Fig. 3a and Extended Data Fig. 7). Genes in pMF01 and pMF03 were placed under the control of the *spoIIE* promoter from *C. sporogenes* ATCC 15579, which is expressed during the late stages of *Clostridium* growth[34], while *baiG* in pMF02 was driven by the strong *fdx* promoter. We conjugated these plasmids sequentially into *C. sporogenes* to yield strain MF001.

When incubated with cholic acid, MF001 produces DCA in a time-dependent manner, in contrast with a control strain that contains only the transporter (*baiG*) (Fig. 3b, c), which does not. Additionally, MF001 converts CDCA to LCA (Extended Data Fig. 6). These data show that the eight genes in the core *bai* cluster (Fig. 1) are sufficient to confer bile acid 7α-dehydroxylation on *C. sporogenes*, although they do not rule out the participation of one or more genes endogenous to *C. sporogenes*.

Identifying branch points in the pathway

To uncover potential branch points for engineering the biosynthesis of non-native pathway products, we constructed a set of strains in which each of the eight genes was individually deleted (Extended Data Fig. 7). We grew these strains with cholic acid and assayed their culture supernatant for the build-up of intermediates (Fig. 3d). Deletion of genes in the oxidative arm of the pathway resulted in the build-up of early pathway intermediates, as expected. Two exceptions were the *baiE* mutant, which produced only cholyl-CoA; and the *baiF*-deficient strain, which generated a small quantity of the final product DCA, suggesting that there might be a compensatory CoA hydrolase or that nonenzymatic hydrolysis of the CoA thioester happens to some extent in vivo.

Intriguingly, the *baiH* mutant accumulates a key intermediate in the reductive arm of the pathway, 3-oxo-4,5-6,7-didehydro-DCA (Fig. 3d), supporting our finding that BaiH catalyses the first reductive step in the pathway. Moreover, strains of *C. sporogenes* expressing BaiG/BaiH and BaiG/BaiCD convert, respectively, 3-oxo-4,5-6,7-didehydro-DCA to 3-oxo-4,5-dehydro-DCA and 3-oxo-4,5-dehydro-DCA to 3-oxo-DCA (Fig. 3e), providing access to intermediates that do not accumulate in a culture of *C. scindens*. Notably, the fully oxidized and partially reduced intermediates are branch points for the production of *allo* (5α) bile acids,

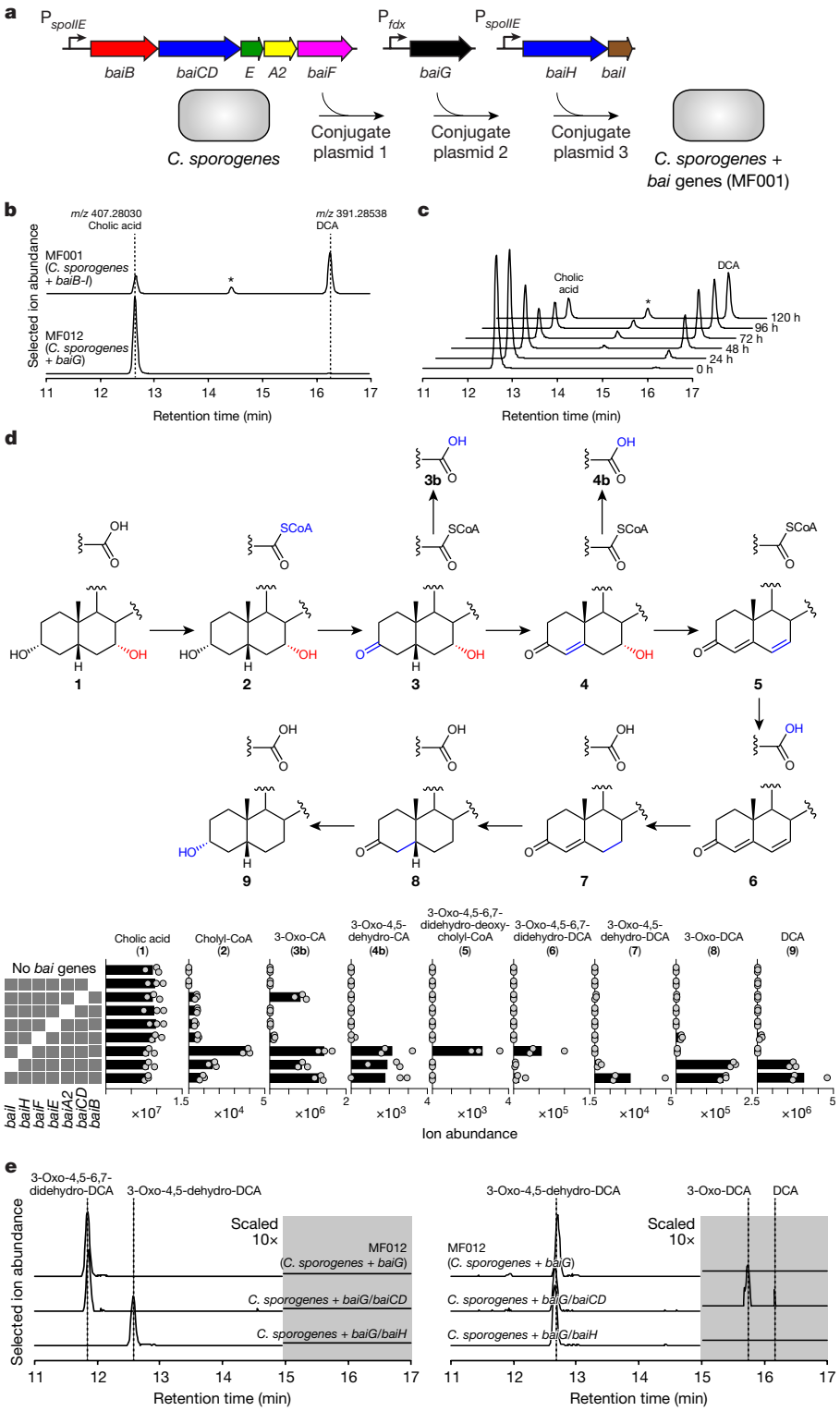


Fig. 3 | Transferring the 7α-dehydroxylation pathway into *C. sporogenes*. **a**, We divided the *bai* operon among three plasmids: *baiB–baiF* in pMTL83153 (plasmid 1, or pMF01), *baiG* in pMTL83353 (plasmid 2, or pMF02) and *baiH–baiI* in pMTL83253 (plasmid 3, or pMF03). These were under the control of the *spoIIE* or *fdx* promoters (P). We successively conjugated pMF01, pMF02 and pMF03 into *C. sporogenes* ATCC 15579 to create MF001. **b**, Combined EICs showing the conversion of cholic acid to DCA by MF001 versus a control strain of *C. sporogenes* harbouring the transporter *baiG* (MF012). The strains were grown with 1 μM cholic acid for 72 h, extracted with acetone, and analysed by LC–MS. The asterisk indicates isoDCA. The experiment was repeated independently three times with similar results. **c**, Combined EICs showing time-dependent conversion of cholic acid to DCA by MF001. The strain was grown with 1 μM cholic acid and aliquots from the indicated time points were analysed as in **b**. The experiment was repeated independently twice with similar results. **d**, Top, simplified proposed pathway for the 7α-dehydroxylation of cholic acid to DCA, including the off-pathway hydrolysis products observed here. Bottom, LC–MS ion abundances for DCA, pathway intermediates and derivatives produced by *C. sporogenes* strains with single gene deletions within the *bai* operon (as indicated at the left). Bars indicate means of three independent biological replicates. **e**, Combined EICs showing the conversion of 3-oxo-4,5-6,7-didehydro-DCA to 3-oxo-4,5-dehydro-DCA by *C. sporogenes* plus *baiG/baiH* (left), and the conversion of 3-oxo-4,5-dehydro-DCA to 3-oxo-DCA by *C. sporogenes* plus *baiG/baiCD* (right). Each strain was cultivated with synthetic 3-oxo-4,5,6,7-didehydro-DCA (left) or 3-oxo-4,5-dehydro-DCA (right) for 72 h and culture extracts were analysed as in **b**. The portions of the traces in the grey boxes have been scaled up tenfold to make it easier to visualize the peaks corresponding to 3-oxo-DCA and DCA. The experiment was repeated independently twice with similar results.

which have important biological activities, including the induction of regulatory T cells[35]. Thus, gaining genetic control over the pathway by expressing it in an alternative gut microbe provides opportunities for the rational and deliberate control of bile acid metabolism and the production of alternative molecules with distinct biological properties.

Colonizing mice with engineered *C. sporogenes*

Finally, we colonized germ-free mice with MF001 to see whether it would confer the production of pathway products on the host. We included two other experimental groups: as a negative control, germ-free mice monocolonized by *C. sporogenes* plus *baiG* (the bile acid transporter); and as a positive control, germ-free mice monocolonized by wild-type *C. scindens*, a native *bai*-operon-containing (7α-dehydroxylating) strain. As shown in Fig. 4, the engineered *C. sporogenes* plus *baiB–I* strain (MF001) conferred production of DCA on the host. The level of production was substantially lower than that observed from *C. scindens*; we suspect that this is because of the need to include additional genes, not yet known, that couple the pathway to a pool of reduced cofactor, increasing flux. Nonetheless, our data

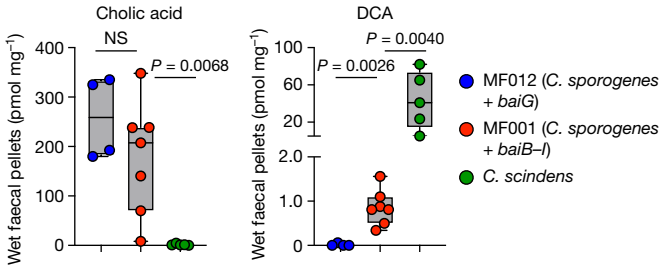


Fig. 4 | In vivo activity of the 7α-dehydroxylation pathway. Germ-free mice were monocolonized with *C. sporogenes* harbouring the transporter *baiG* (the MF012 strain), *C. sporogenes* harbouring *baiB–baiI* (the MF001 strain) or the native 7α-dehydroxylating strain *C. scindens*. Faecal pellets were obtained on day 6 and analysed by LC–MS to determine the quantity of the host-derived substrate cholic acid and the pathway product DCA. *P*-values were determined by a two-tailed *t*-test; NS, not significant. Box and whisker plots show median values, the 25th–75th percentiles, and the range for $n = 4–7$ independent biological replicates.

provide an important starting point for efforts to study and engineer the bile acid pool.

Engineering pathways from the microbiome

Our results reveal the complete bile acid 7α-dehydroxylation pathway, bringing it closer to the level of knowledge we have about endogenous human metabolic pathways. Key features of the pathway might serve as a model for other pathways that produce high-abundance metabolites in the gut (see Supplementary Discussion and Extended Data Fig. 8).

The gut microbiome harbours hundreds of pathways, many of which may modulate host biology, but so far only a few have been the target of engineering[36,37]. This stands in contrast to natural product pathways from terrestrial and marine microorganisms and plants, which are commonly expressed in heterologous hosts[38,39] and engineered to generate non-native products[40]. Two technology gaps need to be overcome in order to make microbiome-derived pathways amenable to engineering: first, we need efficient strategies to identify pathways for known metabolites and small-molecule products of orphan gene clusters, and second, we need tools for transferring pathways into bacterial hosts native to the gut and manipulating them to produce novel molecules. The work described here is a starting point for these efforts. If it can be generalized to other *Clostridia* species, it could lead to a set of tools for de-orphaning, heterologously expressing, and engineering pathways from the microbiome.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2396-4.

1. Buffie, C. G. et al. Precision microbiome reconstitution restores bile acid mediated resistance to *Clostridium difficile*. *Nature* **517**, 205–208 (2015).
2. Yoshimoto, S. et al. Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. *Nature* **499**, 97–101 (2013); corrigendum 506, 396 (2014).
3. Duboc, H., Taché, Y. & Hofmann, A. F. The bile acid TGR5 membrane receptor: from basic research to clinical application. *Dig. Liver Dis.* **46**, 302–312 (2014).
4. Arab, J. P., Karpen, S. J., Dawson, P. A., Arrese, M. & Trauner, M. Bile acids and nonalcoholic fatty liver disease: molecular insights and therapeutic perspectives. *Hepatology* **65**, 350–362 (2017).
5. Nicholson, J. K. et al. Host-gut microbiota metabolic interactions. *Science* **336**, 1262–1267 (2012).
6. Lee, W.-J. & Hase, K. Gut microbiota-generated metabolites in animal health and disease. *Nat. Chem. Biol.* **10**, 416–424 (2014).
7. Koppel, N., Maini Rekdal, V. & Balskus, E. P. Chemical transformation of xenobiotics by the human gut microbiota. *Science* **356**, eaag2770 (2017).
8. Donia, M. S. & Fischbach, M. A. Small molecules from the human microbiota. *Science* **349**, 1254766 (2015).
9. Patel, K. P., Luo, F. J.-G., Plummer, N. S., Hostetter, T. H. & Meyer, T. W. The production of p-cresol sulfate and indoxyl sulfate in vegetarians versus omnivores. *Clin. J. Am. Soc. Nephrol.* **7**, 982–988 (2012).
10. Bouatra, S. et al. The human urine metabolome. *PLoS ONE* **8**, e73076 (2013).
11. Furusawa, Y. et al. Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* **504**, 446–450 (2013); erratum 506, 254 (2014).
12. Maslowski, K. M. et al. Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. *Nature* **461**, 1282–1286 (2009).
13. Smith, P. M. et al. The microbial metabolites, short-chain fatty acids, regulate colonic T_{reg} cell homeostasis. *Science* **341**, 569–573 (2013).
14. Wells, J. E., Berr, F., Thomas, L. A., Dowling, R. H. & Hylemon, P. B. Isolation and characterization of cholic acid 7α-dehydroxylating fecal bacteria from cholesterol gallstone patients. *J. Hepatol.* **32**, 4–10 (2000).
15. Ridlon, J. M., Kang, D.-J. & Hylemon, P. B. Bile salt biotransformations by human intestinal bacteria. *J. Lipid Res.* **47**, 241–259 (2006).
16. Hamilton, J. P. et al. Human cecal bile acids: concentration and spectrum. *Am. J. Physiol. Gastrointest. Liver Physiol.* **293**, G256–G263 (2007).
17. de Aguiar Vallim, T. Q., Tarling, E. J. & Edwards, P. A. Pleiotropic roles of bile acids in metabolism. *Cell Metab.* **17**, 657–669 (2013).
18. Wahlström, A., Sayin, S. I., Marschall, H.-U. & Bäckhed, F. Intestinal crosstalk between bile acids and microbiota and its impact on host metabolism. *Cell Metab.* **24**, 41–50 (2016).
19. Brestoff, J. R. & Artis, D. Commensal bacteria at the interface of host metabolism and the immune system. *Nat. Immunol.* **14**, 676–684 (2013).
20. White, B. A., Lipsky, R. L., Fricke, R. J. & Hylemon, P. B. Bile acid induction specificity of 7α-dehydroxylase activity in an intestinal Eubacterium species. *Steroids* **35**, 103–109 (1980).
21. Ridlon, J. M., Harris, S. C., Bhowmik, S., Kang, D.-J. & Hylemon, P. B. Consequences of bile salt biotransformations by intestinal bacteria. *Gut Microbes* **7**, 22–39 (2016).
22. Mallonee, D. H., Lijewski, M. A. & Hylemon, P. B. Expression in *Escherichia coli* and characterization of a bile acid-inducible 3α-hydroxysteroid dehydrogenase from *Eubacterium* sp. strain VPI 12708. *Curr. Microbiol.* **30**, 259–263 (1995).
23. Mallonee, D. H., Adams, J. L. & Hylemon, P. B. The bile acid-inducible *baiB* gene from *Eubacterium* sp. strain VPI 12708 encodes a bile acid-coenzyme A ligase. *J. Bacteriol.* **174**, 2065–2071 (1992).
24. Bhowmik, S. et al. Structural and functional characterization of BaiA, an enzyme involved in secondary bile acid synthesis in human gut microbe. *Proteins* **82**, 216–229 (2014).
25. Kang, D.-J., Ridlon, J. M., Moore, D. R., II, Barnes, S. & Hylemon, P. B. *Clostridium scindens baiCD* and *baiH* genes encode stereo-specific 7α/7β-hydroxy-3-oxo-Δ^4-cholenoic acid oxidoreductases. *Biochim. Biophys. Acta* **1781**, 16–25 (2008).
26. Dawson, J. A., Mallonee, D. H., Björkhem, I. & Hylemon, P. B. Expression and characterization of a C24 bile acid 7 alpha-dehydratase from Eubacterium sp. strain VPI 12708 in *Escherichia coli*. *J. Lipid Res.* **37**, 1258–1267 (1996).
27. Ye, H. Q., Mallonee, D. H., Wells, J. E., Björkhem, I. & Hylemon, P. B. The bile acid-inducible baiF gene from Eubacterium sp. strain VPI 12708 encodes a bile acid-coenzyme A hydrolase. *J. Lipid Res.* **40**, 17–23 (1999).
28. Harris, S. C. et al. Identification of a gene encoding a flavoprotein involved in bile acid metabolism by the human gut bacterium *Clostridium scindens* ATCC 35704. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* **1863**, 276–283 (2018).
29. González-Pajuelo, M. et al. Metabolic engineering of *Clostridium acetobutylicum* for the industrial production of 1,3-propanediol from glycerol. *Metab. Eng.* **7**, 329–336 (2005).
30. Higashide, W., Li, Y., Yang, Y. & Liao, J. C. Metabolic engineering of *Clostridium cellulolyticum* for production of isobutanol from cellulose. *Appl. Environ. Microbiol.* **77**, 2727–2733 (2011).
31. Kovács, K. et al. Secretion and assembly of functional mini-cellulosomes from synthetic chromosomal operons in *Clostridium acetobutylicum* ATCC 824. *Biotechnol. Biofuels* **6**, 117 (2013).
32. Mingardon, F., Chanal, A., Tardif, C. & Fierobe, H.-P. The issue of secretion in heterologous expression of *Clostridium cellulolyticum* cellulase-encoding genes in *Clostridium acetobutylicum* ATCC 824. *Appl. Environ. Microbiol.* **77**, 2831–2838 (2011).
33. Heap, J. T., Pennington, O. J., Cartman, S. T. & Minton, N. P. A modular system for *Clostridium* shuttle plasmids. *J. Microbiol. Methods* **78**, 79–85 (2009).
34. Wang, Y. et al. Bacterial genome editing with CRISPR–Cas9: deletion, integration, single nucleotide modification, and desirable "clean" mutant selection in *Clostridium beijerinckii* as an example. *ACS Synth. Biol.* **5**, 721–732 (2016).
35. Hang, S. et al. Bile acid metabolites control T_H17 and T_{reg} cell differentiation. *Nature* **576**, 143–148 (2019).
36. Sheridan, P. O. et al. Heterologous gene expression in the human gut bacteria *Eubacterium rectale* and *Roseburia inulinivorans* by means of conjugative plasmids. *Anaerobe* **59**, 131–140 (2019).
37. Hao, T. et al. An anaerobic bacterium host system for heterologous expression of natural product biosynthetic gene clusters. *Nat. Commun.* **10**, 3665 (2019).
38. Huo, L. et al. Heterologous expression of bacterial natural product biosynthetic pathways. *Nat. Prod. Rep.* **36**, 1412–1436 (2019).
39. Keasling, J. D. Manufacturing molecules through metabolic engineering. *Science* **330**, 1355–1358 (2010).
40. Pickens, L. B., Tang, Y. & Chooi, Y.-H. Metabolic engineering for the production of natural products. *Annu. Rev. Chem. Biomol. Eng.* **2**, 211–236 (2011).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Bacterial strains, culture conditions and bile acids

C. scindens VPI 12708 and *C. sporogenes* ATCC 15579 were obtained from the Japan Collection of Microorganisms (JCM) and the American Type Culture Collection (ATCC), respectively. Engineered *C. sporogenes* strains used here are shown in Supplementary Table 3. They were cultured in TYG (3% *w*/*v* tryptone, 2% *w*/*v* yeast extract, 0.1% *w*/*v* sodium thioglycolate) broth at 37 °C in an anaerobic chamber from Coy Laboratories. *E. coli* CA434 (HB101/pRK24) was cultured at 37 °C in LB broth supplemented with 12 μg ml^{-1} tetracycline and 100 μg ml^{-1} carbenicillin. In addition, 20 μg ml^{-1} chloramphenicol, 100 μg ml^{-1} spectinomycin or 250 μg ml^{-1} erythromycin was used for the selection of series of plasmids of pMTL83153, pMTL83353 or pMTL83253 respectively. Plasmids used here are shown in Supplementary Table 2. Cholic acid (**1**), chenodeoxycholic acid, deoxycholic acid (**9**) and lithocholic acid were purchased from Sigma-Aldrich. 3-Oxo-cholic acid (**3b**) and 3-oxo-deoxycholic acid (**8**) were purchased from Steraloids. 3-Oxo-4,5-6,7-didehydro-DCA (**6**) and 3-oxo-4,5-dehydro-DCA (**7**) were synthesized using reported procedures[41]. Structural assignments for the remaining pathway intermediates and derivatives shown in Figs. 2, 3 are provisional, and were made on the basis of mass spectra, retention times and comparison to chemically related standards.

Cloning of the *bai* operon

All amplification by polymerase chain reaction (PCR) was conducted using PrimeSTAR Max DNA polymerase (Takara Bio) according to the manufacturer's instructions. Sequences of primers for target genes and cloning vectors are in Supplementary Table 4. For the heterologous expression of *bai* genes under the *fdx* promoter, pMTL vectors were amplified with primers 1 and 2. For the expression of *bai* genes under the *spoIIE* promoter, pMTL vectors harbouring the *spoIIE* promoter were constructed first. pMTL vectors were amplified with primers 1 and 3 to remove the *fdx* promoter, and the *spoIIE* promoter region, which is the 277-base-pair sequence upstream of CLOSPO_01065, was amplified with primers 4 and 5. Then these two PCR fragments were assembled by overlap PCR. The target gene sequences were amplified with primer pairs shown in Supplementary Table 4. PCR fragments were assembled with the amplified fragments of vectors using a Gibson assembly kit (New England Bio Labs). *E. coli* Stbl4 competent cells (Invitrogen) were transformed with the assembled plasmids by electroporation and transformants were confirmed by PCR. Positive clones containing the assembled plasmids were cultivated, with plasmids obtained by miniprep and verified by sequencing.

Heterologous expression in *C. sporogenes*

Bacterial cultures were incubated in a Coy anaerobic chamber under an atmosphere consisting of 10% CO_2, 5% H_2 and 85% N_2. Growth media were prereduced by overnight preincubation in the anaerobic chamber. For the heterologous expression experiments, plasmids were transferred into *C. sporogenes* by conjugation using *E. coli* CA434, which was electroporated with the individual plasmids and recovered overnight in selective media. We collected 1 ml of overnight culture from the resultant transformants. The cell pellet was washed with phosphate-buffered saline (PBS) to remove residual antibiotics and resuspended with 200 μl of an overnight culture of *C. sporogenes* in anaerobic chamber. Eight drops of 25 μl of the suspension were pipetted on a TYG agar plate without antibiotics and the plate was incubated anaerobically at 37 °C for 2 days. The bacterial biomass was scraped up and resuspended in 300 μl of PBS. The whole cell suspension was then plated on TYG agar plates supplemented with 250 μg ml^{-1} D-cycloserine and appropriate antibiotics (15 μg ml^{-1} thiamphenicol for pMTL83153, 500 μg ml^{-1} spectinomycin for pMTL83353 or 5 μg ml^{-1} erythromycin for pMTL83253). After a few days, antibiotic-resistant colonies were picked and restreaked on agar containing the same antibiotic. The resulting clones were confirmed by PCR amplification using appropriate primers (Supplementary Table 4). Multiple plasmids were introduced sequentially, using the same procedure.

Extraction of metabolites

Engineered strains were cultured anaerobically in TYG medium supplemented with appropriate antibiotics from frozen glycerol stocks. We inoculated 10 μl of the overnight culture in 1 ml of TYG medium supplemented with appropriate antibiotics and 1 μM substrate. After 72 h, unless otherwise noted, the culture was extracted with 20% acetone and centrifuged. The supernatant was analysed by LC–MS.

LC–MS analysis of metabolite extracts

Metabolite extracts were analysed using an Agilent 1290 LC system coupled to an Agilent 6530 quadrupole time-of-flight (QTOF) mass spectrometer with a 1.7 μm, 2.1 mm × 100 mm Kinetex C18 column (Phenomenex). Water with 0.05% formic acid (A) and acetone with 0.05% formic acid (B) were used as the mobile phase at a flow rate of 0.35 ml min^{-1} over a 32-min gradient: 0–1 min, 25% B; 1–25 min, 25–75% B; 25–26 min, 75–100% B; 26–30 min, 100% B; 30–32 min 75–25% B. All data were collected in negative-ion mode.

For detection of CoA conjugates and flavin cofactors, a 1.8 μm, 2.1 mm × 50 mm ZORBAX SB-C18 column (Agilent Technologies) and water with 10 mM ammonium acetate pH 9.0 (A) and acetonitrile (B) was used. A flow rate of 0.3 ml min^{-1} was used over the 17-min gradient: 0–2 min, 15% B; 2–14 min, 15–50% B; 14–14.1 min 50–95% B, 14.1–17 min, 85% B. All data were collected in positive-ion mode.

Cloning of *bai* operon genes

To increase the probability of assembling a complete *bai* operon, we cloned the genes encoding *baiB*, *baiA2*, *baiCD*, *baiE*, *baiF*, and *baiH* from *C. scindens* VPI12708, *Clostridium hylemonae* and *Clostridium hiranonis* using the primers in Supplementary Table 5 and the KOD Xtreme Hot Start PCR kit (Millipore) according to the manufacturer's protocol. Each PCR-amplified gene contains ligation-independent cloning (LIC) sites that are complementary to the pSGC vector. PCR products were purified with the Agencourt Ampure XP PCR clean-up kit (Beckman Coulter) according to the manufacturer's protocol. The pSGC vector was prepared for LIC by linearization with the restriction enzyme *Bsa*I. LIC sites were installed by adding T4 DNA polymerase (NEB) to 10 μg of linearized plasmid in a 50 μl reaction containing 2.5 mM GTP, 1× NEB buffer 2, and 1× bovine serum albumin (BSA) for 1 h at 22 °C. T4 DNA polymerase was heat-inactivated by incubation at 75 °C for 20 min. PCR products (in a volume of 2 μl) were treated with T4 DNA polymerase in a 10 μl reaction containing 2.5 mM CTP, 1× NEB buffer 2 and 1× BSA for 1 h at 22 °C. T4 DNA polymerase was heat-inactivated by incubation at 75 °C for 20 min. The LIC reaction was assembled by mixing 15 ng of digested vector DNA with roughly 40 ng of digested PCR product; the reaction mixture was then incubated at 22 °C for 10 min. A 30 μl aliquot of DH10B cells (NEB) was transformed with 2 μl of the LIC reaction mixture using standard bacterial transformation protocols. This cloning procedure adds a His_6 tag to the amino terminus of each protein with the following sequence: MHHHHHHSSGVDLGTENLYFQS. All final constructs were sequence verified (Genescript).

Expression and purification of BaiH and BaiCD

BL-21(DE3) cells containing the pPH151 plasmid were transformed with the pSGC plasmid containing either *BaiCD* or *BaiH*. The transformants were selected on an LB/agar plate containing 50 μg ml^{-1} kanamycin and 34 μg ml^{-1} chloramphenicol. A single colony was used to inoculate 20 ml of LB overnight culture containing the above antibiotics. The

overnight culture was used to inoculate 2 l of Studier's autoinduction media (ZYP-5052 supplemented with 1 mM flavin mononucleotide and 200 μM $FeCl_3$) housed in a 2 l Pyrex media bottle. Cultures were grown with constant aeration using a sparging stone attached to a pressurized, 0.22-μm filtered air source, all in a water bath maintained at 37 °C. After 5 h, aeration was stopped and the culture was placed in an ice bath for 1 h. The culture was returned to a 22 °C water bath and light aeration was resumed. After 5 min, cysteine was added to a final concentration of 600 μM. The culture was grown at 22 °C for roughly 20 h before being harvested by centrifugation at 10,000*g*. Cell pellets were flash frozen and stored in liquid N_2 until purification. All subsequent steps were carried out in an MBraun anaerobic chamber maintained at less than 0.1 ppm oxygen (MBraun, Stratham, NH). Plastics were brought into the chamber and allowed to sit for two weeks before use. All solvents and buffer stocks were degassed by sparging with argon gas for 4 h before being taken into the chamber.

In a typical purification, roughly 30 g of BaiCD or BaiH cell paste was resuspended in 30 ml of lysis buffer containing 50 mM HEPES, pH 7.5, 300 mM KCl, 4 mM imidazole, 10 mM 2-mercaptoethanol (BME), 10% glycerol, 1 mM flavin mononucleotide (FMN), 1 mM flavin adenine dinucleotide (FAD) and 1% Triton-X305. The resuspension was subjected to 50 rounds of sonic disruption (80% output, 3-s pulse on, 12-s pulse off) at 4 °C. The lysate was cleared by centrifugation at 4 °C for 1 h at 15,000*g*. The supernatant was loaded with an ÄKTA express fast protein liquid chromatography (FPLC) system onto a 5 ml fast-flow HisTrap column (GE Healthcare Life Sciences) equilibrated in lysis buffer lacking FMN, FAD and Triton-X305. The column was washed with 10 column volumes of lysis buffer before elution with 5 ml of buffer containing 50 mM HEPES, pH 7.5, 300 mM KCl, 300 mM imidazole, 10 mM BME and 10% glycerol. The fractions containing protein, based on absorbance at 280 nm, were pooled and reconstituted with iron and sulfur as described[42]. The reconstituted proteins were then passed over a HiPrep 16/60 Sephacryl S-200 HR column equilibrated in 20 mM HEPES, pH 7.5, 300 mM KCl, 5 mM dithiothreitol (DTT) and 10% glycerol. The proteins were concentrated to roughly 1 ml with a Vivaspin 20 concentrator (Sartorius Stedium Biotech). The protein concentration was estimated by absorbance at 280 nm (A_{280}) using the extinction coefficient calculated on the basis of its corresponding amino-acid sequence. The presence of FAD and FMN was confirmed by precipitating purified enzyme with sulfuric acid and analysing the resulting supernatant by LC–MS, comparing analytes to authentic standards of FAD and FMN.

Purification of BaiB, BaiA2, BaiE and BaiF

BL-21(DE3) cells containing the pRIL plasmid were transformed with plasmids containing BaiB, BaiA2, BaiE or BaiF. Each transformant was selected on an LB/agar plate containing 50 μg ml^{-1} kanamycin and 34 μg ml^{-1} chloramphenicol. A single colony was used to inoculate 20 ml of LB overnight culture containing the above antibiotics. The overnight culture was used to inoculate 2 l of Studier's autoinduction media (ZYP-5052) housed in a 2l Pyrex media bottle. Cultures were grown with constant aeration using a sparging stone attached to a pressurized, 0.22-μm filtered air source in a water bath at 37 °C. After 5 h, aeration was stopped and the culture was placed in an ice bath for 1 h. The culture was returned to a 22 °C water bath and light aeration was resumed. The culture was grown at 22 °C for roughly 20 h before being harvested by centrifugation at 10,000*g*. Cell pellets were flash frozen and stored in liquid N_2 until purification. All subsequent steps were carried out in an MBraun anaerobic chamber maintained at less than 0.1 ppm oxygen as above with minor modifications. Briefly, in a typical purification, roughly 30–40 g of cell paste was resuspended in 30–40 ml of lysis buffer containing 50 mM HEPES, pH 7.5, 300 mM KCl, 4 mM imidazole, 10 mM BME, 10% glycerol and 1% Triton-X305. The resuspension was subjected to 50 rounds of sonic disruption (80% output, 3-s pulse on, 12-s pulse off) at 4 °C. The lysate was cleared by centrifugation at 4 °C for 1 h at 15,000*g*. The supernatant was loaded with an ÄKTA express FPLC system onto a 5 mL fast-flow HisTrap column (GE Healthcare Life Sciences) equilibrated in lysis buffer lacking Triton-X305. The column was washed with 10 column volumes of lysis buffer before elution with 5 ml of buffer containing 50 mM HEPES, pH 7.5, 300 mM KCl, 300 mM imidazole, 10 mM BME and 10% glycerol. The fractions containing protein, based on absorbance at 280 nm, were pooled and immediately passed over a HiPrep 16/60 Sephacryl S-200 HR column equilibrated in 20 mM HEPES, pH 7.5, 300 mM KCl, 5 mM DTT and 10% glycerol. The proteins were concentrated to roughly 1 ml with a Vivaspin 20 concentrator (Sartorius Stedium Biotech). The protein concentration was estimated by A_{280} using the extinction coefficient calculated on the basis of its corresponding amino-acid sequence.

In vitro reconstitution of bile acid pathway

Six assays each contained 50 mM HEPES pH 7.5, 50 mM KCl, 200 μM NAD, 100 μM CoA and 200 μM ATP. In addition, each assay contained 0.1 mM of between one and six of the following enzymes: BaiB from *C. scindens*, BaiA2 from *C. scindens*, BaiCD from *C. hiranonis*, BaiE from *C. hiranonis*, BaiF from *C. hylemonae* and BaiH from *C. scindens*. All reactions were initiated with the addition of cholic acid and incubated at 22 °C for 30 min before being quenched by addition of an equal volume of 100% acetone. Each assay was performed in triplicate. Product formation was monitored by LC–MS as described above.

Bile acid pathway reconstitution kinetics

To determine the rate of DCA production by the in vitro pathway, we carried out assays with 50 mM HEPES pH 7.5, 50 mM KCl, 200 μM NAD, 100 μM CoA and 200 μM ATP, plus 0.1 mM of each of BaiB from *C. scindens*, BaiA2 from *C. scindens*, BaiCD from *C. hiranonis*, BaiE from *C. hiranonis*, BaiF from *C. hylemonae* and BaiH from *C. scindens*. Reactions were initiated by adding cholic acid and incubated at 22 °C. Samples of the reaction were removed and mixed with an equal volume of 100 mM H_2SO_4 at the designated times. Each assay was performed in triplicate. Product formation was monitored by LC–MS as above.

K_M assay for BaiCD

Kinetic parameters for BaiCD from *C. hiranonis* were determined in assays that contained 0.45 μM enzyme, 50 mM HEPES pH 7.5, 50 mM KCl and 500 μM NADH. Reactions mixtures were incubated for 5 min at 22 °C before being initiated with 3-oxo-4,5-dehydro-deoxycholic acid. Concentrations of substrate were varied between 3.91 μM and 500 μM. We removed 20 μl of samples and mixed them with an equal volume of 100 mM H_2SO_4 to stop the reaction. Product formation was determined by LC–MS as above. Reactions were performed in triplicate and data were fit to the Michaelis–Menten equation by the least-squares method.

K_M assay for BaiH

Kinetic parameters for BaiH from *C. scindens* were determined in assays that contained 0.45 μM enzyme, 50 mM HEPES pH 7.5, 50 mM KCl and 500 μM NADH. Reactions mixtures were incubated for 5 min at 22 °C before being initiated with 3-oxo-4,5,6,7-didehydro-deoxycholic acid. Substrate concentrations were varied between 0.78 μM and 100 μM. We removed 20 μl of samples and mixed them with an equal volume of 100 mM H_2SO_4 to stop the reaction. Product formation was determined by LC–MS as above. Reactions were performed in triplicate and data were fit to the Michaelis–Menten equation by the least-squares method.

Colonization of germ-free mice

Engineered *C. sporogenes* strains or *C. scindens* ATCC 15579 were cultured anaerobically in TYG medium (supplemented with appropriate antibiotics) or BHI medium from frozen glycerol stocks. Of the overnight cultures, 50 μl were used to inoculate 5 ml of fresh growth medium containing 50 μM cholic acid. After 48 h, bacterial cells were pelleted by centrifugation, washed twice with PBS, resuspended in 25% glycerol solution, and stored at −80 °C. Germ-free C57BL/6 mice

(male, roughly 8 weeks of age, $n = 6$ or 7 per group) were obtained from Taconic Biosciences (Hudson, NY), and colonies were maintained in gnotobiotic isolators in accordance with the Administrative Panel on Laboratory Care (APLAC) of the Stanford Institutional Animal Care and Use Committee (IACUC). The mice were maintained on a Teklad Custom Diet (TD.180755, with 0.5% cholic acid). Mice were gavaged once daily for three successive days using the following procedure: glycerol stocks of engineered strains of *C. sporogenes* or *C. scindens* ATCC 15579 were thawed to room temperature, and mice were inoculated by oral gavage with roughly 200 µl of the thawed glycerol stock (about 1×10^7 colony-forming units). Six days after colonization, faecal pellets were collected and mice were euthanized humanely by CO_2 asphyxiation. Caecal contents were also collected, snap-frozen in liquid nitrogen, and stored at −80 °C.

Murine faecal samples (roughly 30 mg) or intestinal contents (roughly 100 mg) were preweighed into a 2 ml screwtop tube containing six 6-mm ceramic beads (Precellys CK28 Lysing Kit). Then, 300 µl or 1 ml of a mixture of ice-cold acetonitrile, methanol and water (4/4/2, *v*/*v*/*v*) was added to each tube, and samples were homogenized by vigorous shaking using a Qiagen Tissue Lyser II at 25/s for 10 min. The resulting homogenates were centrifuged for 15 min at 14,000*g* at 4 °C, and 100 µl of the supernatant was combined with 100 µl of an internal standard solution (2 µM d4-cholic acid in H_2O). The resulting mixtures were filtered through a Durapore PVDF 0.22-µm membrane using Ultrafree centrifugal filters (Millipore, UFC30GV00), and 5 µl was analysed by LC–MS as described above in the Methods section 'LC–MS analysis of metabolite extracts'.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Mass spectrometry data that support our findings have been deposited in MassIVE (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp) under accession code MSV000085048. Source data are provided with this paper.

41. Leppik, R. A. Improved synthesis of 3-keto, 4-ene-3-keto, and 4,6-diene-3-keto bile acids. *Steroids* **41**, 475–484 (1983).
42. Lanz, N. D. RlmN and AtsB as models for the overproduction and characterization of radical SAM proteins. *Methods Enzymol.* **516**, 125–152 (2012).

Acknowledgements We thank C. T. Walsh, D. Dodd, C. O'Loughlin and members of the Fischbach and Almo laboratories for helpful comments on the manuscript. This work was supported by National Institutes of Health (NIH) grants DP1 DK113598 (to M.A.F.), R01 DK110174 (to M.A.F.), P01 HL147823 (to M.A.F.), P01 GM118303-01 (to S.C.A.), U54 GM093342 (to S.C.A.), U54 GM094662 (to S.C.A.) and DP2 HD101401-01 (to C.G.); the Chan–Zuckerberg Biohub (to M.A.F.); a Howard Hughes Medical Institute (HHMI)–Simons Faculty Scholars Award (to M.A.F.); an Investigators in the Pathogenesis of Infectious Disease Award from the Burroughs Wellcome Foundation (to M.A.F.); and the Price Family Foundation (to S.C.A.).

Author contributions M.F., T.L.G., S.C.A. and M.A.F. conceived and designed the experiments. M.F. developed the system for gene-cluster expression in *Clostridium*, and M.F., C.G. and Y.V. performed the bacterial genetics experiments. T.L.G. expressed and purified enzymes and set up biochemical reconstitution experiments. M.F. analysed the data from biochemical and microbiological experiments by LC–MS. M.E.M. and L.C.B. synthesized bile acid intermediates. M.W. and S.H. performed and analysed mouse experiments. M.F., T.L.G., M.W., S.C.A. and M.A.F. analysed data and wrote the manuscript. All authors discussed the results and commented on the manuscript.

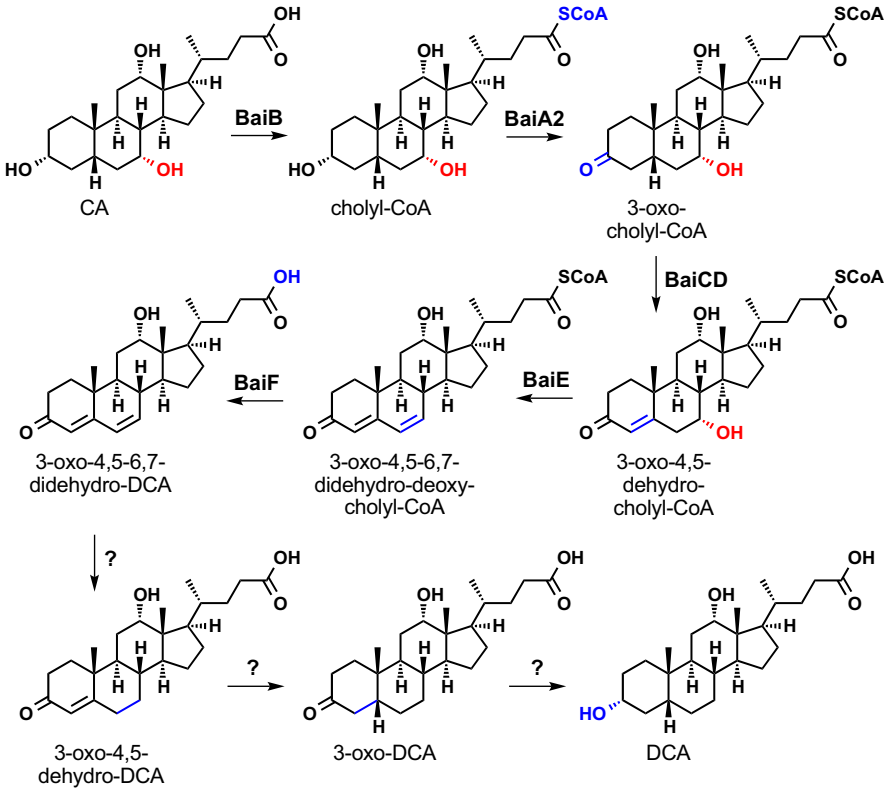
Competing interests M.A.F. is a co-founder and director of Federation Bio.

Additional information

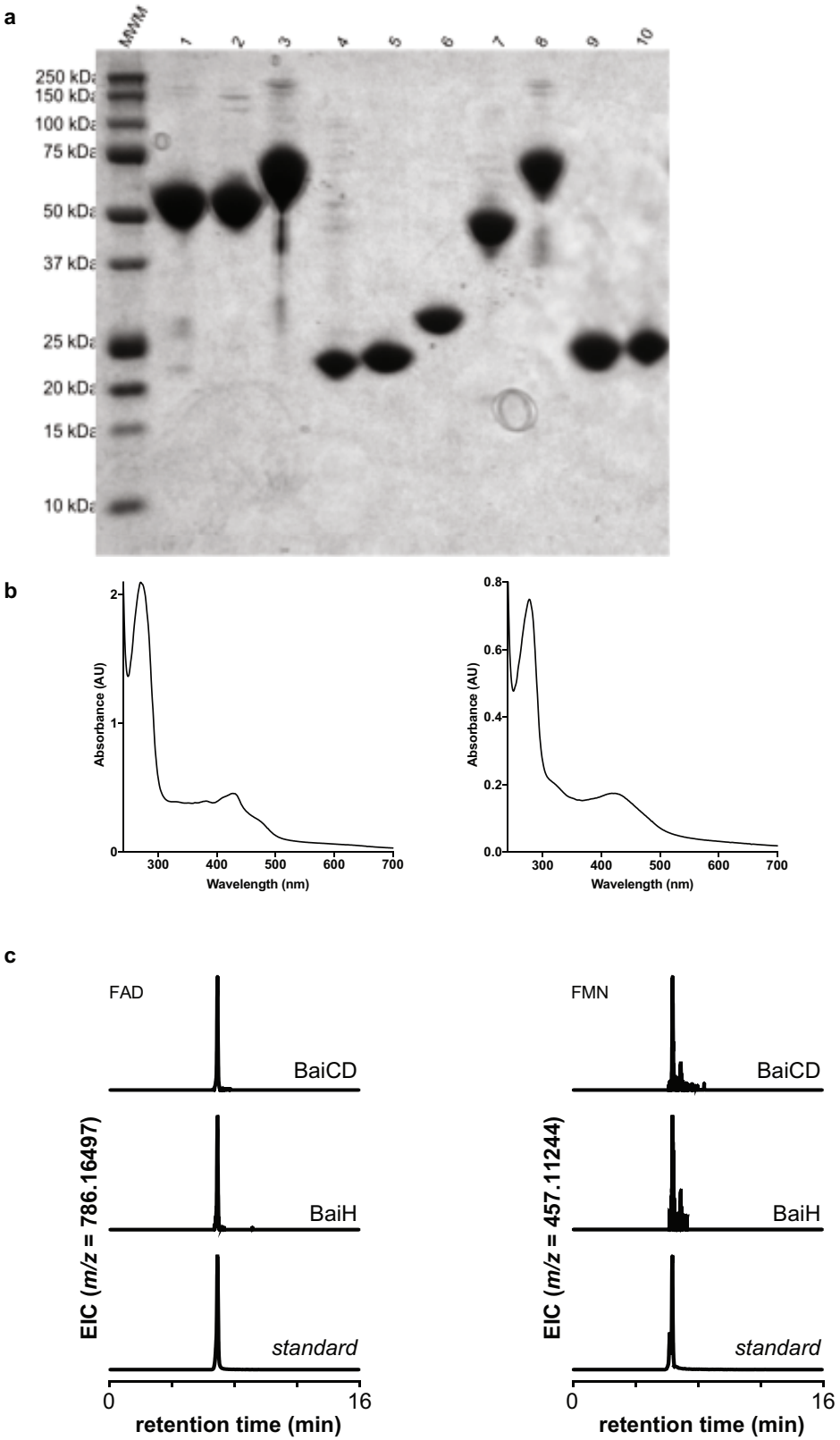
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2396-4.

Correspondence and requests for materials should be addressed to S.C.A. or M.A.F.

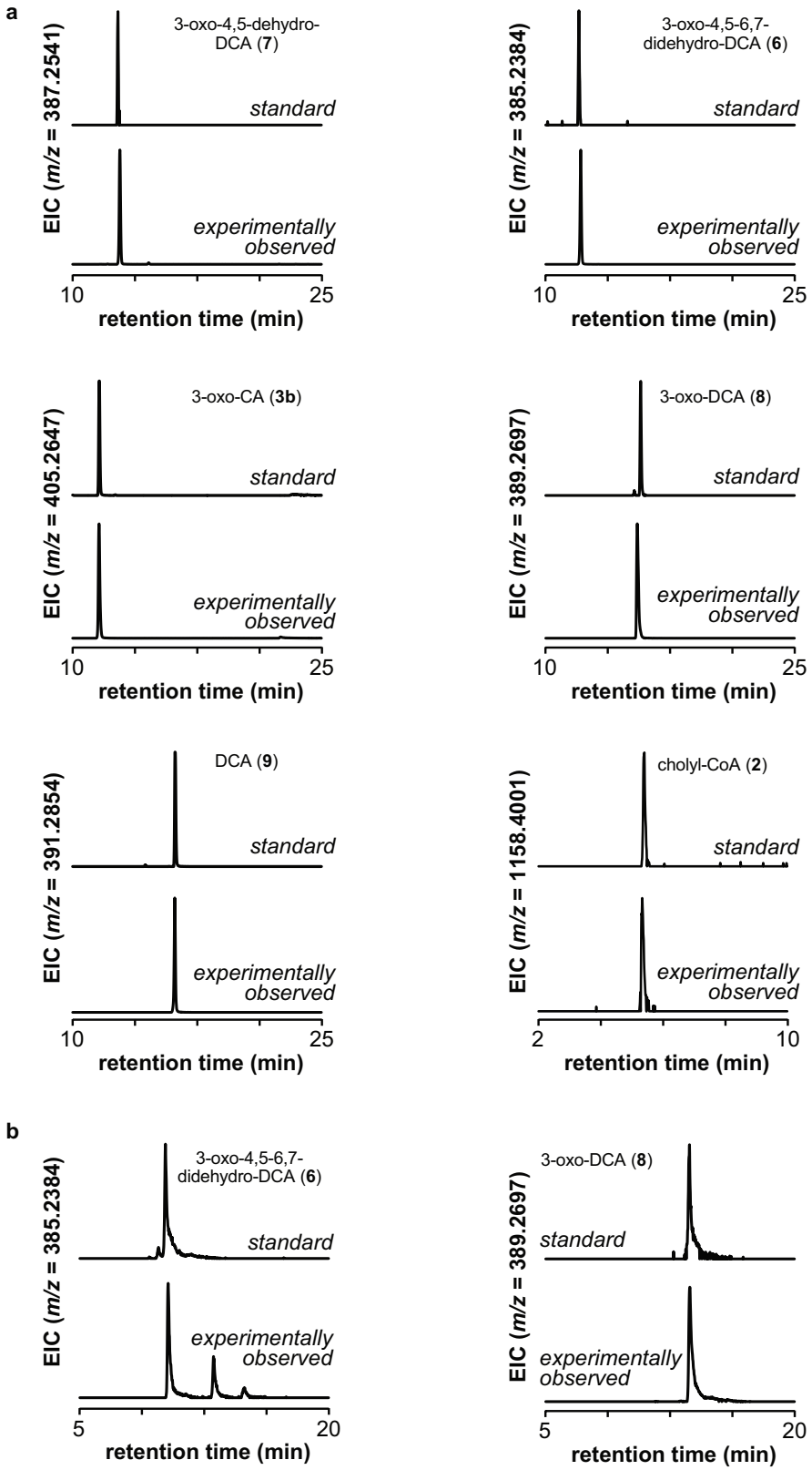
Peer review information *Nature* thanks Pieter Dorrestein and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



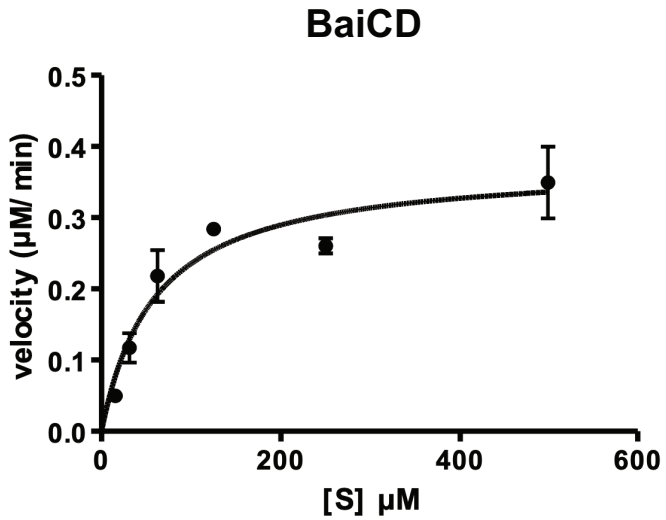
Extended Data Fig. 1 | Previously proposed pathway for 7α-dehydroxylation of cholic acid in *C. scindens* VPI 12708. See main text for details and a summary of the previous literature. CA, cholic acid.

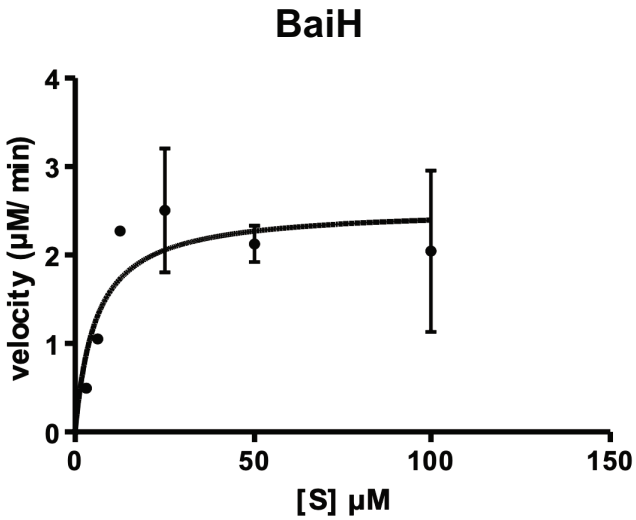


Extended Data Fig. 2 | Purification of recombinant Bai proteins. a, SDS–PAGE analysis of purified Bai proteins after Ni-affinity and size-exclusion purification, visualized by Coomassie blue staining. The image was generated using a Bio-Rad Gel Doc Universal Hood II Molecular Imager. MWM, molecular weight marker; 1, BaiB from *C. scindens*; 2, BaiB from *C. hylemonae*; 3, BaiCD from *C. hiranonis*; 4, BaiE from *C. scindens*; 5, BaiE from *C. hiranonis*; 6, BaiA2 from *C. scindens*; 7, BaiF from *C. hylemonae*; 8, BaiH from *C. scindens*; 9, BaiI from *C. scindens*; 10, BaiI from *C. hiranonis*. **b**, Ultraviolet–visible spectra of BaiCD from *C. hiranonis* (24 μM, left) and BaiH from *C. scindens* (13 μM, right). Features at 370 nm and 450 nm are indicative of flavin bound to BaiCD and BaiH, and are partially obscured by the presence of a [4Fe–4S] cluster, which has broad absorbance between 300 nm and 700 nm. **c**, The presence of FMN and FAD is confirmed by mass spectrometry. Experiments in **a**–**c** were repeated independent twice, with similar results.



Extended Data Fig. 3 | Bile acid standards. a, For each compound in the study for which we have an authentic standard, we show an EIC of the authentic standard and the experimentally observed compound. Because the data shown here were collected from samples run at different times, a drift in retention time may be responsible for the peak pairs that do not have identical retention times. **b**, We observed a drift in retention time in the LC–MS data collected for the experiment shown in Fig. 2c. For two representative compounds from that data set, we show an EIC of the experimentally observed compound and an authentic standard run contemporaneously, showing that the retention times remain consistent with our peak assignments.

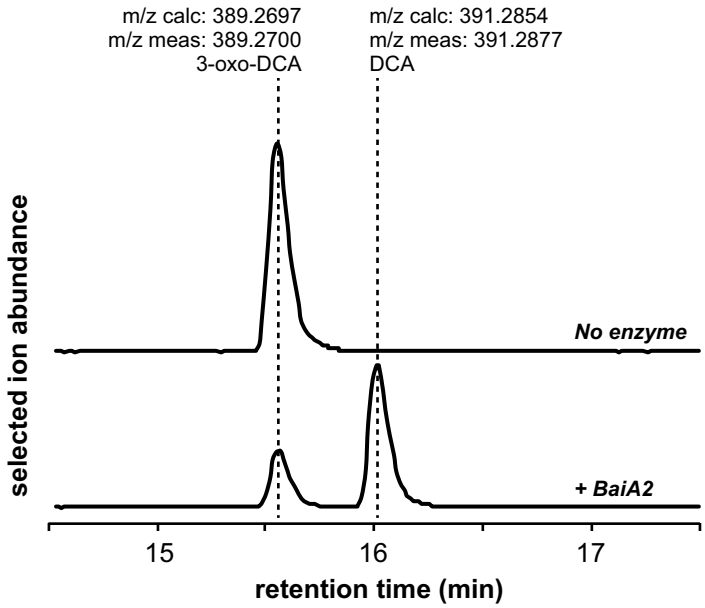




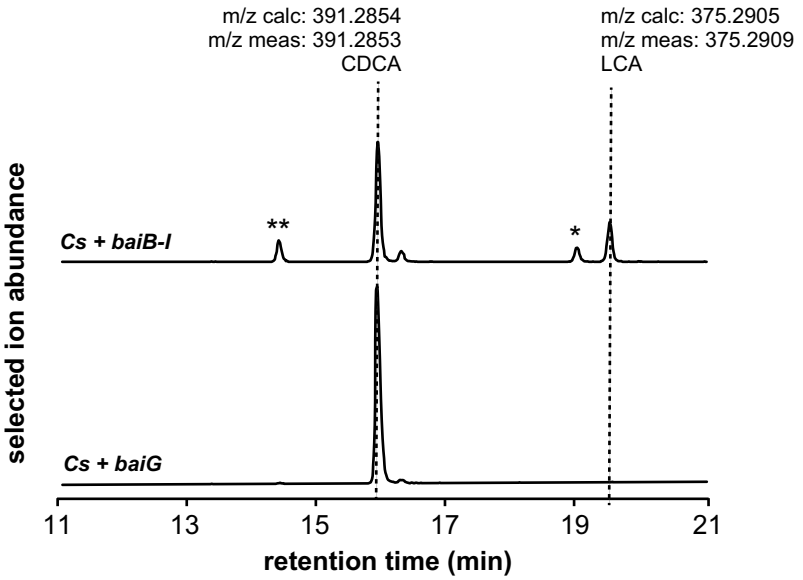
Km 59.7 +/- 13.2 μM
Vmax 0.38 +/- 0.03 μM/min
Kcat 0.84 +/- 0.06 min^{-1}

Km 5.81 +/- 3.88 μM
Vmax 2.53 +/- 0.42 μM/min
Kcat 5.63 +/- 0.93 min^{-1}

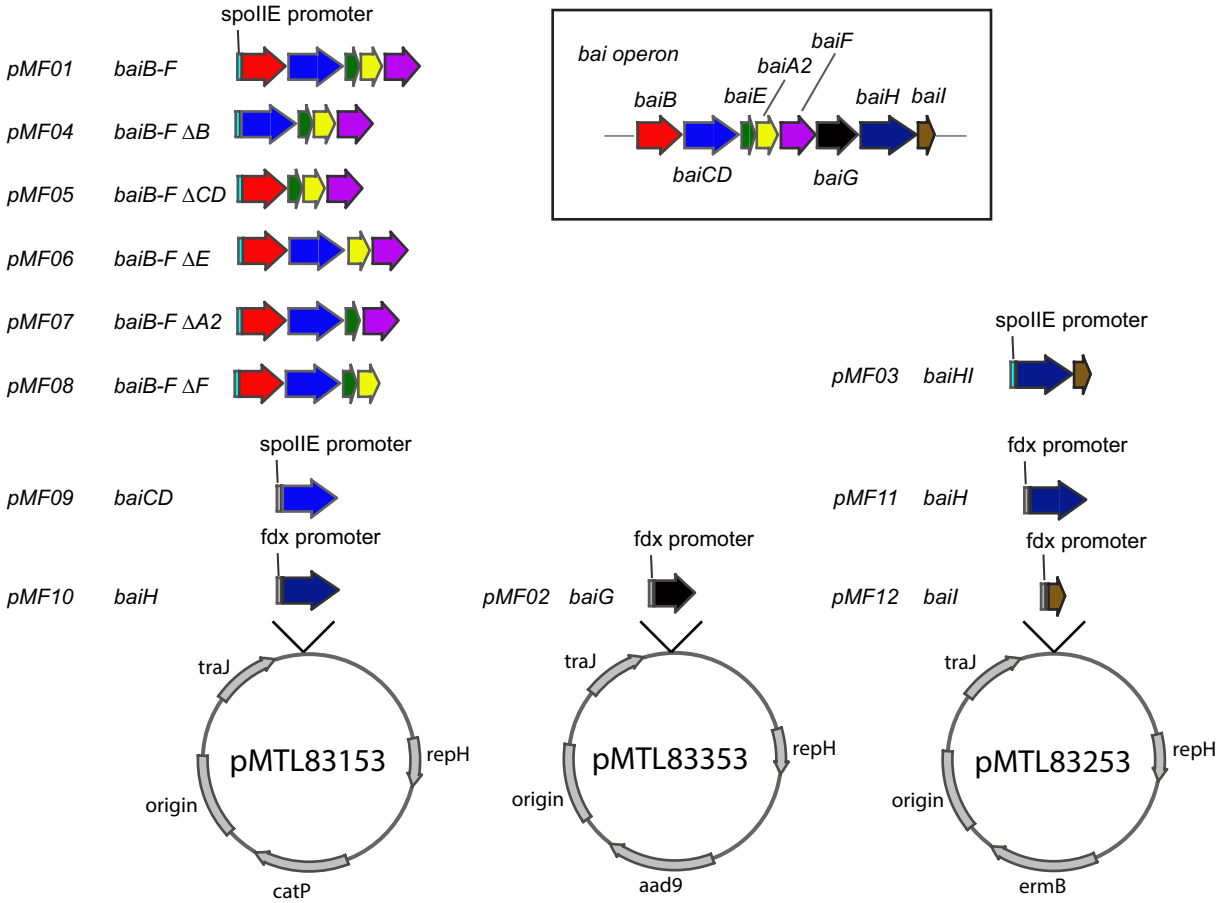
Extended Data Fig. 4 | Kinetic parameters for BaiCD and BaiH. a, Michaelis–Menten analysis of the conversion of 3-oxo-4,5-dehydro-DCA to 3-oxo-DCA by BaiCD. Reaction mixtures contained 0.45 μM BaiCD and 1 mM NADH, with the substrate concentration varying between 15 μM and 500 μM. **b**, Michaelis–Menten analysis of the conversion of 3-oxo-4,5,6,7-didehydro-DCA to 3-oxo-4,5-dehydro-DCA by BaiH. Reaction mixtures contained 0.45 μM BaiH and 1 mM NADH, with the substrate concentration varying between 3 μM and 100 μM. Data indicate the average product level ± 1 s.d. (three biological replicates).



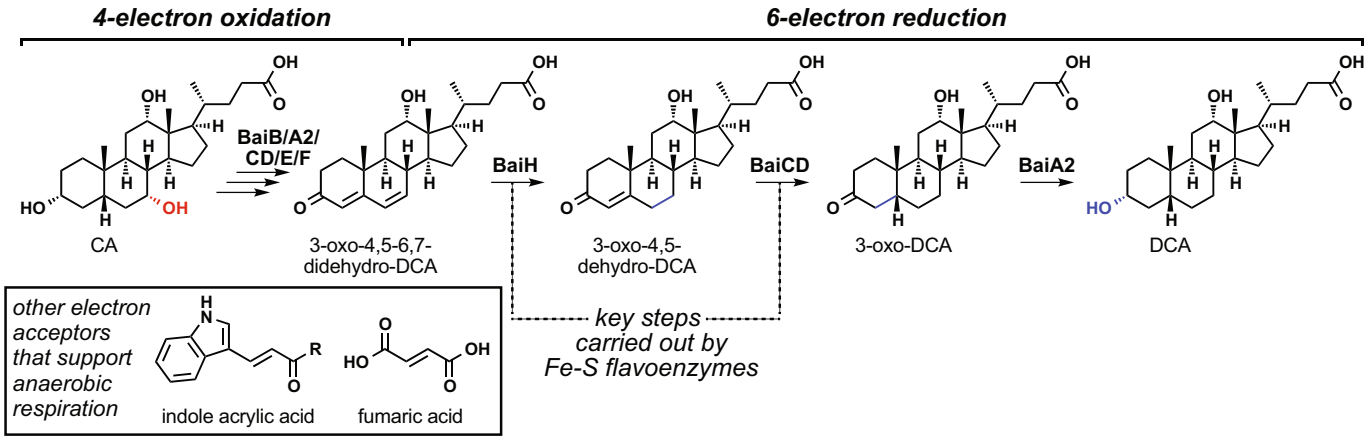
Extended Data Fig. 5 | Biochemical analysis of 3-oxo-DCA reduction by BaiA2. Combined EICs showing the conversion of 3-oxo-DCA to DCA by recombinant BaiA2. This experiment was performed once.



Extended Data Fig. 6 | 7α-dehydroxylation of CDCA in vivo. Combined EICs showing the conversion of CDCA to LCA by a *C. sporogenes* strain harbouring the complete *bai* operon on three plasmids (MF001) versus a control strain of *C. sporogenes* harbouring the transporter *baiG* (MF012). The strains were cultivated with 1 μM cholic acid for 72 h; an acetone extract of the culture supernatant was analysed by high-performance LC (HPLC)/MS. The single asterisk indicates isoLCA; the peak indicated by the double asterisk is provisionally assigned as isoCDCA. This experiment was performed once.



Extended Data Fig. 7 | Constructs for expressing the *bai* operon and portions thereof in *C. sporogenes*. Each of the plasmids has replication origins (origin and repH) for *E. coli* and *Clostridium*, the *traJ* gene to enable conjugal plasmid transfer, and an antibiotic-resistance gene (*catP*, *aad9* or *ermB*). The *bai* genes were introduced into these plasmids under the control of the *fdx* or *spoIIE* promoter. For the genetic analysis of *baiCD* and *baiH* function, pMTL83153-based plasmids were used.



Extended Data Fig. 8 | Metabolic logic of the 7α-dehydroxylation pathway. Highly oxidized metabolic intermediates as anaerobic electron acceptors. In the first half of the 7α-dehydroxylation pathway, two successive two-electron oxidations set up a vinylogous dehydration of the 7-hydroxyl, yielding the highly oxidized intermediate 3-oxo-4,5-6,7-didehydro-DCA. In the second half of the pathway, three successive two-electron reductions reduce this molecule to DCA, resulting in a net two-electron reduction. The first two of these reductions are carried out by Fe–S flavoenzymes, which comprise a suite of four cofactors that enable them to convert two-electron inputs to a one-electron manifold. The previously proposed pathway is shown in Extended Data Fig. 1.

nature research

Corresponding author(s): Michael Fischbach, Steven Almo

Last updated by author(s): Jan 26, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Agilent MassHunter
Data analysis	Agilent MassHunter, GraphPad Prism 7

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

There are no restrictions on data availability. Mass spectrometry data are publicly available in the MassIVE repository.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. In most experiments, a sample size of three was chosen since this is the minimum number required to perform tests of statistical significance. In the mouse experiment, 4-7 mice were used based on past experience with the degree of variability in measuring microbiome-derived metabolites in mouse tissues.
Data exclusions	No data were excluded from the manuscript. In some figure panels, a representative example is shown (e.g., an individual set of chromatograms) in order to highlight a given result.
Replication	Reproducibility was verified by performing independent biological replicates. All attempts at replication were successful.
Randomization	Samples were not randomized. They were run consecutively on the LC-MS; instrument drift within each sample set was minimal.
Blinding	Researchers were not blinded during these experiments. Data were analyzed using software that calculates metabolite levels in an objective manner, so blinding is less relevant in this setting.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	We used germ-free C57/BL6 mice, male, 8 +/- 2 weeks of age.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve field-collected samples.
Ethics oversight	The study protocol was approved by A-PLAC, the Stanford IACUC.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Deciphering human macrophage development at single-cell resolution

https://doi.org/10.1038/s41586-020-2316-7

Received: 15 June 2019

Accepted: 11 March 2020

Published online: 20 May 2020

Check for updates

Zhilei Bian[1,2,3,17], Yandong Gong[4,17], Tao Huang[4,17], Christopher Z. W. Lee[5,6,17], Lihong Bian[7,17], Zhijie Bai[4,17], Hui Shi[4,17], Yang Zeng[8], Chen Liu[4], Jian He[4], Jie Zhou[8], Xianlong Li[4], Zongcheng Li[8], Yanli Ni[8], Chunyu Ma[7], Lei Cui[9], Rui Zhang[10,11], Jerry K. Y. Chan[12,13,14], Lai Guan Ng[5], Yu Lan[1,2]✉, Florent Ginhoux[5,6,15,16]✉ & Bing Liu[1,4,8]✉

Macrophages are the first cells of the nascent immune system to emerge during embryonic development. In mice, embryonic macrophages infiltrate developing organs, where they differentiate symbiotically into tissue-resident macrophages (TRMs)[1]. However, our understanding of the origins and specialization of macrophages in human embryos is limited. Here we isolated $CD45^+$ haematopoietic cells from human embryos at Carnegie stages 11 to 23 and subjected them to transcriptomic profiling by single-cell RNA sequencing, followed by functional characterization of a population of $CD45^+CD34^+CD44^+$ yolk sac-derived myeloid-biased progenitors (YSMPs) by single-cell culture. We also mapped macrophage heterogeneity across multiple anatomical sites and identified diverse subsets, including various types of embryonic TRM (in the head, liver, lung and skin). We further traced the specification trajectories of TRMs from either yolk sac-derived primitive macrophages or YSMP-derived embryonic liver monocytes using both transcriptomic and developmental staging information, with a focus on microglia. Finally, we evaluated the molecular similarities between embryonic TRMs and their adult counterparts. Our data represent a comprehensive characterization of the spatiotemporal dynamics of early macrophage development during human embryogenesis, providing a reference for future studies of the development and function of human TRMs.

Although macrophages are best known for their immune functions, there is accumulating evidence that they have additional roles, including regulating the haematopoietic microenvironment, influencing metabolism, mediating tissue repair, and overseeing the maturation of embryonic tissue[1]. These diverse roles are mirrored by their varied ontogenic pathways, with some TRMs arising from adult bone marrow-derived haematopoietic stem cells (HSCs) and some instead being specified early in development from embryonic haematopoietic precursors[2].

Mammalian embryonic haematopoiesis is a complex process, involving multiple temporally overlapping programs[3–5]. Fate-mapping studies in mice have shown that macrophages develop much earlier during embryonic haematopoiesis than do functional HSCs[6–8], with the earliest progenitors emerging in the yolk sac in at least two distinct waves: an early c-Myb-independent primitive wave arising at embryonic day (E) 7.5, and a later c-Myb-dependent erythromyeloid progenitor (EMP) wave arising at E8.25[9–12]. The progenitors in the first wave differentiate in situ into macrophages that migrate into the brain rudiment, being the major source of microglia[7]. The second wave progenitors, initially defined as EMPs, differentiate in situ into macrophages and also migrate into the fetal liver, giving rise to several lineages, including monocytes that will migrate into diverse developing tissues and begin differentiating into TRMs before birth[13,14].

Although the characteristics and development of embryonic macrophages in mice have been characterized in detail, they are not well understood in humans. We know that the first functional HSCs appear in the aorta–gonad–mesonephros region at Carnegie stage (CS) 14, and slightly later in the yolk sac at CS16, whereas the first HSCs in the liver are detected at CS17[15]. However, further progress has been hampered by the difficulty of obtaining early human embryos for study, and because

[1]Key Laboratory for Regenerative Medicine of Ministry of Education, Institute of Hematology, School of Medicine, Jinan University, Guangzhou, China. [2]Guangzhou Regenerative Medicine and Health-Guangdong Laboratory (GRMH-GDL), Guangzhou, China. [3]Integrated Chinese and Western Medicine Postdoctoral Research Station, Jinan University, Guangzhou, China. [4]State Key Laboratory of Proteomics, Academy of Military Medical Sciences, Academy of Military Sciences, Beijing, China. [5]Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), BIOPOLIS, Singapore, Singapore. [6]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore. [7]State Department of Gynecology, Fifth Medical Center of Chinese PLA General Hospital, Beijing, China. [8]State Key Laboratory of Experimental Hematology, Fifth Medical Center of Chinese PLA General Hospital, Beijing, China. [9]Laboratory of Hematologic Diseases, Beijing Pediatric Research Institute, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China. [10]Beijing Key Laboratory of Pediatric Hematology Oncology, Beijing Children's Hospital, Capital Medical University, Beijing, China. [11]Hematology Oncology Center, National Center for Children's Health, Beijing Children's Hospital, Capital Medical University, Beijing, China. [12]Department of Reproductive Medicine, KK Women's and Children's Hospital, Singapore, Singapore. [13]Department of Obstetrics & Gynaecology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [14]OBGYN-Academic Clinical Program, Duke-NUS, Duke-NUS Medical School, Singapore, Singapore. [15]Shanghai Institute of Immunology, Shanghai JiaoTong University School of Medicine, Shanghai, China. [16]Translational Immunology Institute, SingHealth Duke-NUS Academic Medical Centre, Singapore, Singapore. [17]These authors contributed equally: Zhilei Bian, Yandong Gong, Tao Huang, Christopher Z. W. Lee, Lihong Bian, Zhijie Bai, Hui Shi. ✉e-mail: rainyblue_1999@126.com; florent_ginhoux@immunol.a-star.edu.sg; bingliu17@yahoo.com

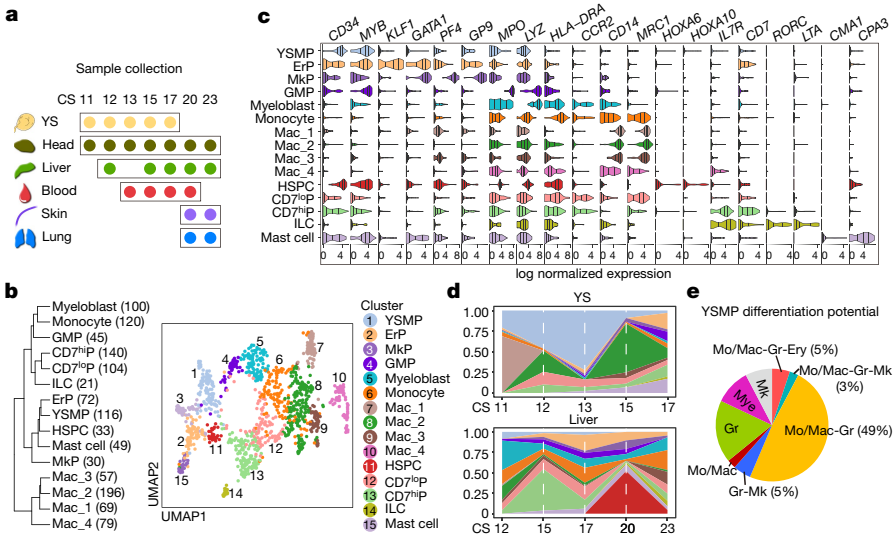


Fig. 1 | Transcriptomic landscape and functional characterization of CD45$^+$ haematopoietic cells in tissues from human embryos. a, Location and Carnegie stage information for samples used for STRT–seq ($n = 8$ biologically independent embryo samples). Details can be found in Supplementary Table 1. **b**, Uniform manifold approximation and projection (UMAP) visualization of all haematopoietic cell clusters identified (right), and hierarchical clustering using Euclidean distance with the number of cells in each cluster ($n = 1,231$ cells total) shown in parentheses (left). Detailed cell information and DEGs can be found in Supplementary Table 2. **c**, Violin plot overview of expression of key lineage-associated genes by the haematopoietic clusters (number of cells in each cluster shown in **b**), with twenty-fifth, fiftieth and seventy-fifth percentiles marked. **d**, Area charts showing changes in the proportions of the haematopoietic clusters (represented by the same colours as in **b**) between CS11 and CS17 for the yolk sac ($n = 5$ biologically independent embryo samples and 238 cells), and CS12 and CS23 for the liver ($n = 6$ biologically independent embryo samples and 354 cells). **e**, Pie chart showing lineage potential of YSMPs defined by single cell culture and subsequent flow cytometry analysis ($n = 39$ clones generated from a CS13 yolk sac). Mo/mac, monocytes/macrophages (CD45$^+$CD33$^+$CD14$^+$); Gr, granulocytes (CD45$^+$CD33$^+$CD66b$^+$); Ery, erythrocytes (CD235a$^+$); Mk, megakaryocytes (CD41a$^+$).

conventional methods are unable to accurately analyse the extremely small number of early embryonic cells isolated.

In this study, we present findings from eight early human embryos obtained after abortion, from which we isolated a range of tissues before carrying out single-cell RNA sequencing (scRNA-seq) on the CD45$^+$ haematopoietic cell populations. This enabled us to study the spatiotemporal distribution and dynamic process of human embryonic haematopoiesis and thereby to answer questions around the origins and specification of human macrophages.

scRNA-seq of human embryo haematopoiesis

To study human macrophage development, we used fluorescence-activated cell sorting (FACS) to isolate CD45$^+$CD235a$^-$ haematopoietic cells from various anatomical sites (yolk sac, head, liver, blood, skin and lung) of eight embryos that had been aborted at multiple Carnegie stages (from CS11 to CS23, equivalent to approximately days 24–56 of development) (Fig. 1a, Supplementary Fig. 1). We then used a modified scRNA-seq approach called single-cell tagged reverse transcription and sequencing (STRT–seq), as previously reported[16–18]. Using high-precision quality-controlled transcriptomic data from 1,231 single cells (Fig. 1b, Extended Data Fig. 1, Supplementary Table 1), we annotated 15 clusters according to the expression of known feature genes (Fig. 1b, c, Supplementary Table 2).

A yolk sac-derived progenitor population and macrophages were among the first CD45$^+$ haematopoietic cells to emerge, appearing in the yolk sac at CS11 (Fig. 1d, Extended Data Fig. 1f, g). Notably, these progenitors showed much weaker transcriptomic erythroid features than mouse EMPs[19] (Extended Data Fig. 1k, l), suggestive of their myeloid-biased nature, and therefore were annotated as yolk sac-derived myeloid-biased progenitors (YSMPs). Both YSMPs and haematopoietic stem and progenitor cells (HSPCs) expressed high levels of *CD34* and *MYB*, but HSPCs were seen only in the liver after CS17, and specifically expressed HOX family transcription factors such as *HOXA6* and *HOXA10* (Fig. 1c, d, Extended Data Fig. 2, Supplementary Table 3). From CS12 onwards, we detected a population of granulocyte–monocyte progenitors (GMPs) in the liver that were characterized by high expression of *MYB*, *MPO* and *LYZ*. This contrasted with CD7^{lo} progenitors that arose at CS12, which expressed both lymphoid-related (*CD7* and *IL7R*) and myeloid (*MRC1* and *CD14*) genes, and with CD7^{hi} progenitors that were identified in the liver from CS15 onwards, which expressed higher levels of *IL7R* and *CD7*. Mast cells, which expressed high levels of *CMA1* and *CPA3*, emerged in the yolk sac at CS13, and then in the liver, blood and skin from CS15 onwards. Innate lymphoid cells (ILCs), which are characterized by *RORC* expression, were detected in the yolk sac from as early as CS17 (Fig. 1c, d).

Notably, two populations (Mac_1 and Mac_4) clustered away from the other groups (Fig. 1b). Mac_1 cells were mainly found in the yolk sac at CS11 (Extended Data Fig. 1f, g) and expressed the yolk sac-related gene *S100A1*[20] as well as high levels of the endothelial marker *CDH5*[21] (Extended Data Fig. 1j, Supplementary Table 2). By contrast, Mac_4 cells were predominantly located in the head, and this was the only population that expressed *SALL1*[22] and gliomedin (*GLDN*)[23], indicating that these cells represent developing microglia (Extended Data Fig. 1j).

We next validated our STRT–seq results by high-throughput 10x Genomics analysis (Extended Data Fig. 3a–f, Supplementary Table 4). First, we re-analysed the yolk sac (YS) cells that had been sampled between CS11 and CS17 by STRT–seq, and found two macrophage populations: YS-Mac_1 cells were found mainly in the CS11 embryo, and YS-Mac_2 cells were found predominantly at CS15. 10x Genomics data from CS11 and CS15 yolk sacs similarly identified two macrophage populations separated stage-wise. We then projected the top ten differentially expressed genes (DEGs) identified by STRT–seq from both the YS-Mac_1 and YS-Mac_2 clusters on to the 10x Genomics data and found similar gene expression patterns. Finally, we identified CD34 and CD44 as putative markers of the YSMP population

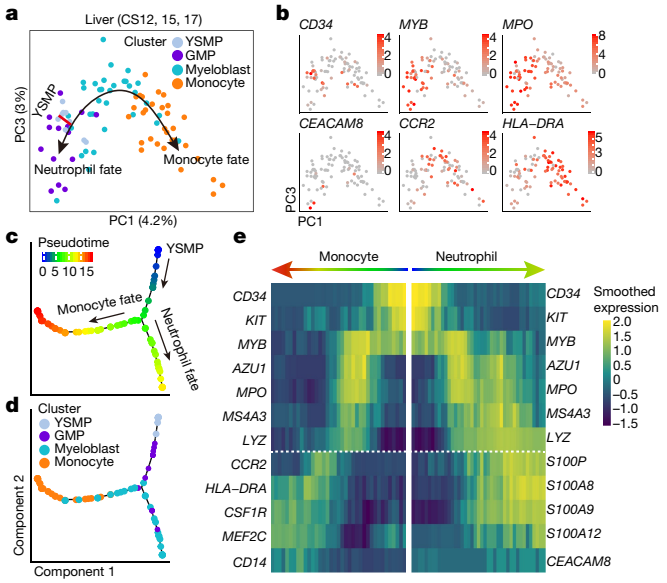


Fig. 2 | Developmental trajectory of YSMPs in human embryonic liver. **a**, PCA of YSMP, GMP, myeloblast and monocyte populations ($n = 88$ cells) sampled from livers at CS12–CS17 ($n = 4$ biologically independent embryo samples) suggests two distinct fates of YSMPs. **b**, Expression levels of the indicated genes projected onto PCA. **c**, **d**, Monocle prediction of YSMP developmental trajectory with pseudotime (**c**) and cluster information (**d**) mapped on. **e**, Heat map showing scaled expression of branching curated genes of monocyte and neutrophil fates ordered by pseudotime.

(Extended Data Fig. 3g, h) and sorted these cells to determine their functional identity.

Bulk cultures of yolk sac cells revealed that haematopoietic progenies were readily detected in the CD45$^+$CD34$^+$CD44$^+$ population (YSMPs), with mostly CD33$^+$ myeloid cells, but not in the CD45$^+$CD34$^-$CD44$^-$ population (Extended Data Fig. 4a–c). Haematopoietic clusters were found in 67 (36%) of 184 YSMP single-cell cultures (Extended Data Fig. 4d). Thirty-nine of these culture wells were then selected randomly and analysed using flow cytometry. Sixty-two per cent (24 wells) had multi-lineage potential, with most of them (22 wells) containing both monocytes/macrophages (CD14$^+$) and granulocytes (CD66b$^+$). Of note, 92% (36 wells) exhibited myeloid potential (CD33$^+$), among which 23 wells showed monocyte/macrophage differentiation, in contrast to only 5% (2 wells) with erythroid potential (Fig. 1e, Extended Data Fig. 4d, e). These data confirmed the myeloid-biased multi-lineage potential of human YSMPs, in line with their transcriptional features (Extended Data Fig. 1k, l).

YSMP development in embryonic liver

To investigate whether human YSMPs could also give rise to HSPC-independent monocytes in vivo, we studied the myeloid-related clusters in liver up to CS17, a stage before transcriptomically defined HSPCs were first detected. Principal component analysis (PCA) suggested that the YSMPs, which expressed *CD34* and *MYB*, were located between the GMP and myeloblast populations (Fig. 2a, b, Extended Data Fig. 5a, b). These cells gradually expressed neutrophilic genes such as *CEACAM8*, while also increasing their expression of the monocyte markers *CCR2* and *HLA-DRA* (Fig. 2b), indicating that YSMPs can differentiate along two distinct paths. Trajectory analysis using Monocle similarly revealed two distinct cell fates arising from YSMPs (Fig. 2c, d, Extended Data Fig. 5c). Based on the gene expression profiles, it appeared likely (although we were unable to capture mature neutrophils in our analysis) that YSMPs could give rise to both monocytes and neutrophils in vivo, similar to our observations in the in vitro functional assay of YSMPs. Comparing the DEGs between the two branches revealed that the cells shared a common signature including expression of *MYB*, *MPO* and *MS4A3*, while the monocyte branch exclusively expressed genes such as *CCR2* and *CD14*, and the neutrophil branch began to express canonical neutrophil markers such as *S100A9* and *CEACAM8* (Fig. 2e, Extended Data Fig. 5d–g, Supplementary Table 5).

Two waves of yolk sac-derived embryonic TRMs

An important question for us to answer was whether the origin and specification of macrophages is similar in humans to those in mice, in which there are two yolk sac-derived waves: a monocyte-independent primitive wave in early yolk sac, and a later fetal liver monocyte-derived wave[14].

To study this, we first re-clustered the identified macrophages and macrophage-related populations alone (Fig. 3a, Extended Data Fig. 6a, b), and annotated them on the basis of their unique expression characteristics, staging information and localization, supporting our annotation by comparison with a curated list of mouse TRM-specific genes[24] (Extended Data Fig. 6c–e). A closer look at the unique DEGs between these populations suggested that these cells may have already initiated their tissue residency programs, with the Liver_Mac population expressing *SPIC*, the Blood_Mac population expressing *CCL13*, and the Lung_Mac and Skin_Mac populations expressing their tissue-related *BMX* and *MMP1* genes, respectively (Fig. 3b, Supplementary Table 6). Hierarchical clustering revealed that Head_Mac3, Head_Mac4 and Liver_Mac were clustered away from the rest of the macrophage groups (Fig. 3c): we hypothesized that this may be because they were more mature TRMs and therefore selected them for further analysis.

The YS_Mac1 group corresponded to the Mac_1 group that we previously identified (Fig. 1c), which uniquely expressed the endothelial gene *CDH5*[21] among macrophages (Fig. 3d) and notably also expressed the red blood cell-related *HBE1* gene. High expression of *HBE1* is a hallmark of yolk sac-derived nucleated primitive erythrocytes[25], which led us to question whether YS_Mac1/Mac_1 might be a related lineage belonging to the early yolk sac-derived primitive macrophage wave. At CS11, this population was found mainly in the yolk sac, although some were also present in the head (Extended Data Fig. 6c–e). Together, our analyses suggest that these cells are yolk sac-derived primitive macrophages, with some of them migrating early to the head as microglial precursors, as in mouse development[7].

After establishing the distinct identity of the YS_Mac1 population, we investigated the developmental trajectory of early human yolk sac-derived embryonic macrophages, considering their dual origins. We analysed macrophage-associated populations between CS11 and CS17, before the appearance of HSPCs, which are likely to arise from a separate definitive lineage. A temporal assessment of the Carnegie staging of these populations revealed that YS_Mac1 cells appear earliest, at CS11, and contribute to the main macrophage populations first, whereas YSMPs gave rise to monocyte-derived macrophages, but only after CS17 (Fig. 3e). The expression of key lineage-defining genes showed a similar pattern (Fig. 3d), with YSMPs expressing higher levels of *MYB* and giving rise to *MPO*-expressing progenitors, consistent with studies on fetal monocytes in zebrafish[11]. Likewise, the YS_Mac1 population had a strong macrophage identity from the start, expressing *CD163* and *MRC1*. Both YSMPs and YS_Mac1 cells expressed *CDH5*, consistent with their yolk sac-derived endothelial origins. Thus, our data suggest that early embryonic macrophage development in humans closely mirrors the processes seen in mice and zebrafish.

Human microglia origin and specification

In mice, microglia primarily arise from the first wave of yolk sac-derived primitive macrophages[7]. To study the origin and specification of

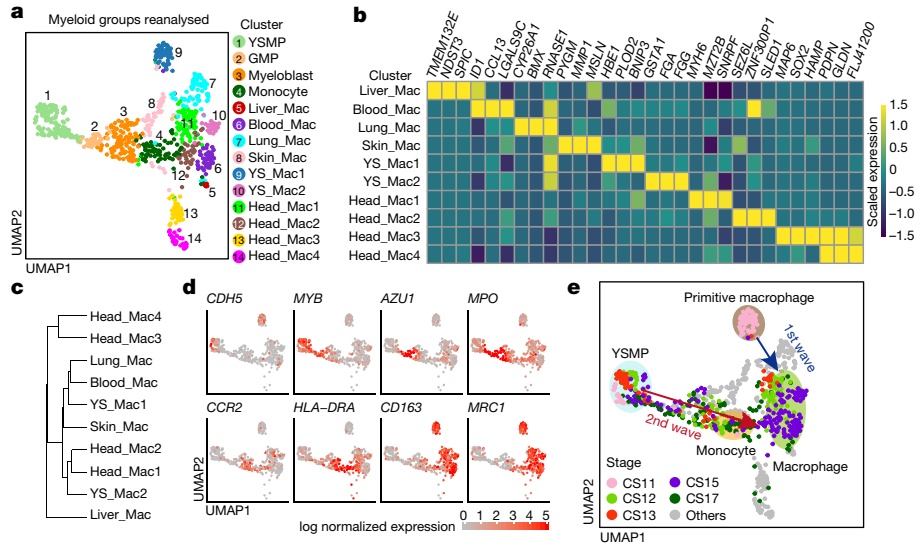


Fig. 3 | Two distinct waves of yolk sac-derived macrophages contribute to TRM populations in human embryos. a, UMAP visualization of myeloid-related cells ($n = 8$ biologically independent embryo samples and 782 cells) with 14 re-clustered populations mapped on. **b**, Heat map showing scaled expression of the top three DEGs between macrophage populations ($n = 450$ cells). DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with *P* value adjusted for multiple testing using Bonferroni correction), and the genes with fold change >1.5 and adjusted $P < 0.05$ were selected (Supplementary Table 6). **c**, Hierarchical clustering based on Euclidean distance of macrophage populations ($n = 450$ cells). **d**, Expression of the indicated genes projected on to UMAP of myeloid-related clusters. **e**, UMAP visualization of myeloid-related clusters with cells from CS11 to CS17 ($n = 6$ biologically independent embryo samples and 268 cells) mapped on, showing sequential appearance of macrophage populations.

microglia in human embryos, we focused on macrophages sampled from head tissues at CS11–CS23. Clustering analysis revealed that these macrophages could be divided into five groups (Head_Mac0–Head_Mac4), with features highly related to their developmental stage. The signature of Head_Mac0 overlapped completely with that of YS_Mac1; together with the observation that the yolk sac was the sole source of macrophages at this time-point, this suggests that YS_Mac1 gives rise to the macrophages found in the head at CS11 (Fig. 4a). A small population of monocytes derived from YSMPs was also seen at CS17 in the head. However, the disparity between their number and those of the other Head_Mac populations, as well as their late appearance, together suggested that they are likely to make only a minor contribution (Extended Data Fig. 7a, b). The sequential emergence of the five head macrophage populations indicated that these populations were part of a continuum of developing cells, and their expression of classical microglia-associated genes including *CX3CR1*, *SALL1* and *SPP1* further supported the idea that these cells were moving towards a microglial fate[22,26] (Fig. 4b).

Next, we studied the dynamics of microglial specification by generating a profile of gene expression changes across the five Head_Mac clusters based on their sequential real-time emergence between CS11 and CS23 (Fig. 4c, Supplementary Table 7). The changes largely followed five patterns. Most genes in patterns 1 and 2 were downregulated and some of them, such as *CD163* and *CD1C*, were associated with immune activity. We also witnessed increased expression of tissue development and neurodevelopmental genes such as *IGF1*[26] and *TMSB4X*[27]. A survey of the transcription factor landscape revealed a similar trend (Fig. 4d), with head macrophages losing expression of inflammatory transcription factors such as *IRF7* and *STAT1*, and gaining microglial identity with expression of *BHLHE41*, *JUN*, *FOSB*, *NR4A1* and *SALL1*[22]; these changes mirror the pattern observed during mouse microglial development.

Finally, we integrated our embryonic data with publicly available data for adult TRMs from the head, liver and lungs, along with our own data for skin sampled from children aged eight and ten. The results confirmed that specification had already occurred in the microglial population in embryos, with the embryonic and adult microglia clustering together (Extended Data Fig. 8, Supplementary Table 8). Likewise, specification towards liver TRM fate had begun, although to a lesser extent than for head macrophages (Extended Data Fig. 7d–g, Supplementary Table 9). This is in contrast with skin TRMs, where there was only minor expression of *CD207* at CS23 (Extended Data Fig. 7h) and slight overlap between the embryonic and paediatric samples in the integrated dataset (Extended Data Fig. 8a, b). Similarly, no commitment had yet occurred in the embryonic lung (Extended Data Figs. 7c, 8a, b, Supplementary Table 8).

Discussion

Much of our current knowledge of embryonic haematopoiesis is based on findings in animal models such as mice or zebrafish. This study paves the way for a wide range of explorations and analyses that were previously difficult to approach owing to the ambiguity surrounding human yolk sac-derived macrophages and their progenitors. Although it is widely acknowledged that mammalian haematopoiesis is highly conserved[3], the characterization of these cells in humans has so far been restricted to either microscopic observations[3,28] or explant experiments[29]. By leveraging the maturation of single-cell sequencing technology and bioinformatics, our analyses shed light on this issue in an unbiased and unsupervised manner, while maintaining tissue site and temporal information. Although it is still difficult to determine with certainty the ontogeny of the various human TRM subsets without the use of fate-mapping tools typically used in mouse models, we have identified two distinct HSC-independent waves of macrophages in humans that correspond to those seen in mice. This is especially important in the clinical context, as macrophages are essential regulators of tissue development and homeostasis[30], and understanding their functions and developmental pathways is key to the diagnosis and treatment of pathologies caused by their dysregulation. The contribution of these HSC-independent waves must be considered when characterizing the subsets of macrophages found in disease, as numerous studies have established that HSC-independent macrophages maintain a distinct transcriptomic and epigenetic identity from their HSC-derived counterparts[31]. Specifically, the human YSMPs that we have transcriptomically and functionally characterized here might correspond to mouse

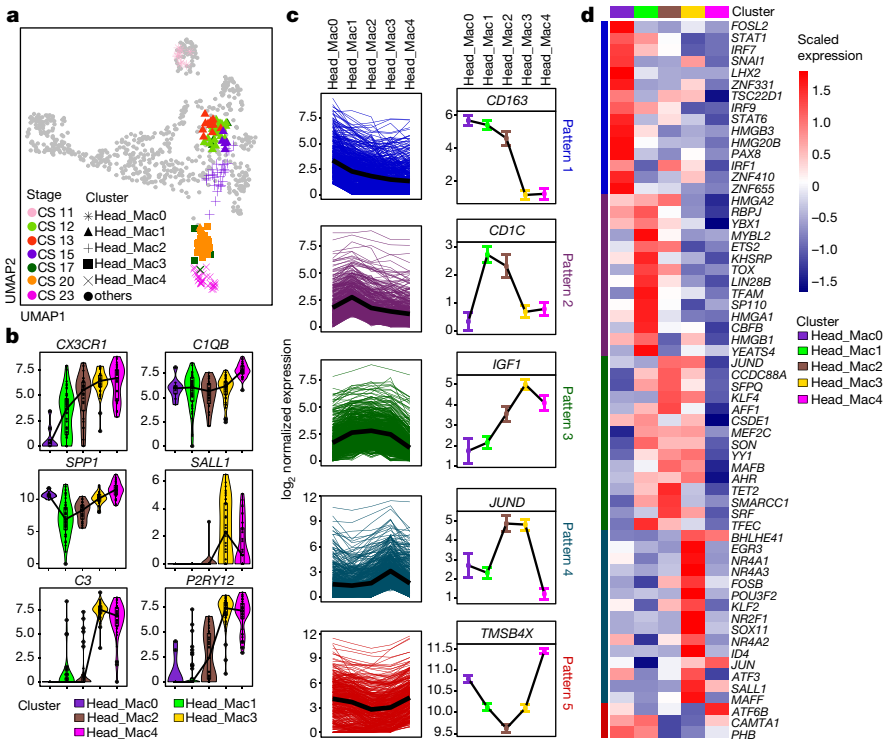


Fig. 4 | Origin and specification of microglia in human embryos. a, UMAP visualization of myeloid-related cells with macrophages in human embryonic head ($n = 8$ biologically independent embryo samples and 155 cells) mapped on. Stage and cluster information indicated by colours and symbols, respectively. **b**, Violin plots showing changes in expression of microglia-related genes between head macrophage clusters. For box plot within each violin plot, centre black lines indicate median values, boxes range from the 25th to 75th percentiles, and whiskers correspond to 1.5× interquartile range (IQR). **c**, Five main patterns revealed by DEGs between head macrophage clusters. Left, expression levels of all pattern genes (coloured lines) and the average expression of each pattern (black line). Complete gene list can be found in Supplementary Table 7. Right, kinetics of expression of a representative gene for each pattern (mean ± s.e.m.). **d**, Heat map showing scaled expression of transcription factors within the pattern genes. Colours to the left indicate the patterns as in **c**.

EMPs, which exhibit similar spatiotemporal dynamics and haematopoietic progenitor characteristics[13]. However, erythroid potential is readily observed in cultures of mouse EMPs[13], consistent with their prominent erythroid properties transcriptomically[19] (Extended Data Fig. 1k, l). These observations warrant further investigation towards their in-depth comparison and the identification of related governing mechanisms.

The embryonic macrophage signature identified in our study also presents an interesting opportunity to re-evaluate our commonly used TRM models. Conventionally, for human studies, macrophages have been generated from blood monocytes via in vitro culture with CSF-1. Comparison of these in vitro monocyte-derived macrophages with the bona fide yolk sac-derived macrophages that initially seed the embryonic tissues might reveal crucial similarities and differences between these cells and allow better experimental design and interpretation of results. Our study also supports the case for a greater understanding of induced pluripotent stem (iPS) cell-derived macrophages and the various TRM systems that have been developed based on these cells[32]. Although a direct comparison between mouse primitive macrophages and iPS cell-derived macrophages has already proven that they are similar to each other and yet distinct from HSC-derived macrophages[33], the same comparison has not, to our knowledge, been carried out for their human counterparts. Our study will contribute to an understanding of the ontogeny of these iPS cell-derived macrophages.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2316-7.

1. Wynn, T. A., Chawla, A. & Pollard, J. W. Macrophage biology in development, homeostasis and disease. *Nature* **496**, 445–455 (2013).
2. Ginhoux, F. & Guilliams, M. Tissue-resident macrophage ontogeny and homeostasis. *Immunity* **44**, 439–449 (2016).
3. Palis, J. et al. Spatial and temporal emergence of high proliferative potential hematopoietic precursors during murine embryogenesis. *Proc. Natl Acad. Sci. USA* **98**, 4528–4533 (2001).
4. Dzierzak, E. & Bigas, A. Blood development: hematopoietic stem cell dependence and independence. *Cell Stem Cell* **22**, 639–651 (2018).
5. Zhou, F. et al. Tracing haematopoietic stem cell formation at single-cell resolution. *Nature* **533**, 487–492 (2016).
6. Samokhvalov, I. M., Samokhvalova, N. I. & Nishikawa, S. Cell tracing shows the contribution of the yolk sac to adult haematopoiesis. *Nature* **446**, 1056–1061 (2007).
7. Ginhoux, F. et al. Fate mapping analysis reveals that adult microglia derive from primitive macrophages. *Science* **330**, 841–845 (2010).
8. Yona, S. et al. Fate mapping reveals origins and dynamics of monocytes and tissue macrophages under homeostasis. *Immunity* **38**, 79–91 (2013).
9. Hoeffel, G. & Ginhoux, F. Ontogeny of tissue-resident macrophages. *Front. Immunol.* **6**, 486 (2015).
10. Frame, J. M., McGrath, K. E. & Palis, J. Erythro-myeloid progenitors: "definitive" hematopoiesis in the conceptus prior to the emergence of hematopoietic stem cells. *Blood Cells Mol. Dis.* **51**, 220–225 (2013).
11. Bertrand, J. Y. et al. Definitive hematopoiesis initiates through a committed erythromyeloid progenitor in the zebrafish embryo. *Development* **134**, 4147–4156 (2007).
12. Ivanovs, A. et al. Human haematopoietic stem cell development: from the embryo to the dish. *Development* **144**, 2323–2337 (2017).
13. McGrath, K. E. et al. Distinct sources of hematopoietic progenitors emerge before HSCs and provide functional blood cells in the mammalian embryo. *Cell Rep.* **11**, 1892–1904 (2015).
14. Hoeffel, G. et al. C-Myb^{+} erythro-myeloid progenitor-derived fetal monocytes give rise to adult tissue-resident macrophages. *Immunity* **42**, 665–678 (2015).
15. Ivanovs, A. et al. Highly potent human hematopoietic stem cells first emerge in the intraembryonic aorta-gonad-mesonephros region. *J. Exp. Med.* **208**, 2417–2427 (2011).

16. Ivanovs, A., Rybtsov, S., Anderson, R. A., Turner, M. L. & Medvinsky, A. Postmenstrual gestational age should be used with care in studies of early human hematopoietic development. *Blood* **121**, 3051–3052 (2013).
17. Zeng, Y. et al. Tracing the first hematopoietic stem cell generation in human embryo by single-cell RNA sequencing. *Cell Res.* **29**, 881–894 (2019).
18. Zeng, Y. et al. Single-cell RNA sequencing resolves spatiotemporal development of pre-thymic lymphoid progenitors and thymus organogenesis in human embryos. *Immunity* **51**, 930–948.e936 (2019).
19. Dong, J. et al. Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. *Genome Biol.* **19**, 31 (2018).
20. Kiewitz, R. et al. S100A1, a new marker for acute myocardial ischemia. *Biochem. Biophys. Res. Commun.* **274**, 865–871 (2000).
21. Costa, G. et al. SOX7 regulates the expression of VE-cadherin in the haemogenic endothelium at the onset of haematopoietic development. *Development* **139**, 1587–1598 (2012).
22. Buttgereit, A. et al. Sall1 is a transcriptional regulator defining microglia identity and function. *Nat. Immunol.* **17**, 1397–1406 (2016).
23. Gosselin, D. et al. An environment-dependent transcriptional network specifies human microglia identity. *Science* **356**, eaal3222 (2017).
24. Mass, E. et al. Specification of tissue-resident macrophages during organogenesis. *Science* **353**, aaf4238 (2016).
25. Peschle, C. et al. Haemoglobin switching in human embryos: asynchrony of ζ→α and ε→γ-globin switches in primitive and definite erythropoietic lineage. *Nature* **313**, 235–238 (1985).
26. Wlodarczyk, A. et al. A novel microglial subset plays a key role in myelinogenesis in developing brain. *EMBO J.* **36**, 3292–3308 (2017).
27. Wirsching, H. G. et al. Thymosin β4 induces folding of the developing optic tectum in the chicken (*Gallus domesticus*). *J. Comp. Neurol.* **520**, 1650–1662 (2012).
28. Enzan, H. Electron microscopic studies of macrophages in early human yolk sacs. *Acta Pathol. Jpn.* **36**, 49–64 (1986).
29. Tavian, M., Robin, C., Coulombel, L. & Péault, B. The human embryo, but not its yolk sac, generates lympho-myeloid stem cells: mapping multipotent hematopoietic cell fate in intraembryonic mesoderm. *Immunity* **15**, 487–495 (2001).
30. Gordon, S. & Plüddemann, A. Tissue macrophages: heterogeneity and functions. *BMC Biol.* **15**, 53 (2017).
31. Lavin, Y. et al. Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. *Cell* **159**, 1312–1326 (2014).
32. Lee, C. Z. W., Kozaki, T. & Ginhoux, F. Studying tissue macrophages in vitro: are iPSC-derived cells the answer? *Nat. Rev. Immunol.* **18**, 716–725 (2018).
33. Takata, K. et al. Induced-pluripotent-stem-cell-derived primitive macrophages provide a platform for modeling tissue-resident macrophage differentiation and function. *Immunity* **47**, 183–198.e186 (2017).

Methods

Human samples and quality control

The embryos were obtained from pregnant women undergoing medical abortions at The Fifth Medical Center of the PLA General Hospital (Beijing, China) the women had provided informed consent. The integrity and morphology of the embryos were evaluated, the somite pairs were counted, and the crown–rump lengths were measured under the microscope to define developmental stage[34]. Paediatric skin samples from children aged eight and ten years were obtained from tissue discarded after elective circumcision at Beijing Children's Hospital, Capital Medical University with parental consent. All protocols were approved by the institutional review boards (approval number: ky-2017-3-5 and IEC-C-008-A08-2018-75) and in accordance with the regulations of the Declaration of Helsinki. All protocols were compliant with the Interim Measures for the Administration of Human Genetic Resources, administered by the Ministry of Science and Technology of China.

Preparation of cell suspensions and FACS

Embryonic tissues (yolk sac, head, lung, skin, and liver) were carefully dissected under the microscope, with the exception of blood, which was collected using a suction pipe directly from the heart. After being washed in PBS three times to remove any residual blood contamination, the yolk sac, head, lung and skin were transferred to pre-warmed digestion medium containing 0.1 g/ml collagenase I (Sigma, C2674, preheated to 37 °C) in RPMI 1640 medium (Gibco,11875093). The samples were enzymatically digested at 37 °C in a humidified incubator for 20–30 min, with the samples being shaken periodically every 5 min until they were digested into a single-cell suspension. The liver was first cut into small pieces using scalpels, and then mechanically dissociated using syringes into a single-cell suspension, before removal of erythrocytes with lysis buffer (BD). The cells were then filtered through a 70-μm cell strainer after the enzymes were neutralized using serum. Child skin samples were first incubated in RPMI 1640 medium with 1.2 U/ml dispase II (Roche) at 4 °C for 8 h before separation of the epidermal layer. After separation, the epidermal layer was cut into pieces and digested with 0.25% trypsin-EDTA (Gibco) containing 0.25 mg/ml DNase I (Sigma DN25) for 30 min at 37 °C in a humidified incubator. After that, the epidermal layer was passed through a 70-μm cell strainer by grinding, and then washed with PBS.

The following antibodies were used to label the embryonic cells: CD45 (BV421, 563879 BD, lot 9066960), CD235a (APC-Cy7, 349116 Biolegend, lot 7355682) and 7-amino-actinomycin D (7-AAD) (PerCP-Cy5.5, 00699350 eBioscience, lot 1910559). Langerhans cells were isolated using CD45 (BV421, 563879 BD, lot 9066960), 7-AAD (PerCP-Cy5.5, 00699350 eBioscience, lot 1910559), CD207 antibodies (PE, 564727 BD, lot 8135683) and CD1a antibodies (APC, 559775 Biolegend, lot 8164562). Cells were sorted using an Aria 2 Flow Cytometer (BD Bioscience). Data were analysed using BD FACSDIVA V8.0.1 and Flowjo (V10).

Haematopoietic progenitor culture in vitro

MS5 stromal cells[29] were seeded into 48-well or 96-well flat-bottom plates with α-MEM (Gibco, 12561-056) supplemented with 20% fetal bovine serum (HyClone, SH30070.03) at a density of 20,000 (48-well) or 10,000 cells (96-well) per well about 24 h before use. One hour before co-culture, stromal medium was replaced with serum-free StemPro 34 (Gibco, 10639011) supplemented with 50 ng/ml hSCF (PeproTech, 300-07-2), 50 ng/ml hFLT3 ligand (PeproTech, 300-19-2), 10 ng/ml hIL-3 (PeproTech, 200-03-2), 10 ng/ml hIL-6 (PeproTech, 200-06-5), 5 ng/ml hIL-11 (PeproTech, 200-11-10), 25 ng/ml hTPO (PeproTech, 300-18-10), 20 ng/ml hGM-CSF (PeproTech, 300-03-5), 3 U/ml EPO (PeproTech, 100-64-10), 1% 2-mercaptoethanol (Gibco, 21985-023), 1% L-glutamine (Gibco, 21051024) and 1% penicillin/streptomycin (Gibco, 15140-122).

Candidate cells were sorted into wells and cultured at 37 °C for 10 or 14 days. The morphologies of haematopoietic clusters were imaged using a Leica camera (A05C872000). At the end of the cultures, all cells including MS5 stromal cells in the selected positive wells were collected and FACS analysis was used for identification of myeloid cells (Mye, $CD45^+CD33^+$), monocytes/macrophages (Mo/Mac, $CD45^+CD33^+CD14^+$), granulocytes (Gr, $CD45^+CD33^+CD66b^+$), erythrocytes (Ery, $CD235a^+$) and megakaryocytes (Mk, $CD41a^+$).

For bulk cultures, YSMP ($CD45^+CD34^+CD44^+$) and negative control ($CD45^+CD34^-CD44^-$) populations were sorted and co-cultured with MS5 for 14 days (100 cells per well). The assays were performed in $n = 3$ biologically independent experiments (using one CS11 and two CS12 yolk sacs, 21 wells in total). For the single cell cultures, 184 single YSMPs from a CS13 yolk sac were sorted and cultured in individual wells. Wells were observed under a light microscope and those with haematopoietic clusters (more than 50 round haematopoietic-like cells) at day 10 were counted as positive wells, which were stochastically chosen for further analyses with FACS to determine the lineage differentiation potential. For each well, 20,000 living cells were recorded, which took around one-third of the total number. The numbers of haematopoietic cells (including $CD45^+$, $CD235a^+$ and $CD41^+$ cells) generated in each well thus ranged from 200 to 4,500 according to the calculation. Lineage potential was ascertained when the number of cells that expressed the given lineage marker (described above) was more than 20.

The following antibodies were used for staining and sorting: anti-CD45 (BV421, 563879 BD, lot 9066960), anti-CD34 (PE, 550761 BD, lot 7129824), anti-CD44 (BV605, 562991 BD, lot 7103609). The following antibodies were used for the identification of haematopoietic clusters and lineages: anti-CD235a (APC-Cy7, 349116 BioLegend, lot B289027), CD235a (Pacific Blue, 306611 BioLegend, lot B224563), anti-CD41a (APC, 17-0419-42 eBioscience, lot 2073742), anti-CD45 (FITC, 11-0459-42 eBioscience, lot 4310016), anti-CD33 (PE, 555450 BD, lot 8074660), anti-CD33 (APC-Cy7, 366614 BioLegend, lot B252646), anti-CD14 (BV786, 563698 BD, lot 8351911) and anti-CD66b (PE, 561650 BD, lot 7264511).

Single-cell RNA-seq library preparation and sequencing

Sequencing libraries were constructed following a modified single-cell tagged reverse transcription (STRT) protocol as previously reported[35–37]. The cells were first judged by morphology under a microscope to assess their condition, with cells in good condition being picked by mouth pipette and directly placed into lysis buffer. The reverse transcription reaction was performed using a sample-specific 25-nucleotide (nt) oligo dT primer containing an 8-nt barcode (TCAGACGTGTGCTCTTCCGATCT-XXXXXXXX-DDDDDDDD-T25, where X represents the sample-specific barcode, and D stands for the unique molecular identifier (UMI)). After reverse transcription and second-strand cDNA synthesis, the cDNAs were amplified by 16 cycles of PCR. The barcoded DNAs were then pooled together and purified using Agencount AMPure XP beads. Biotinylated pre-indexed primers were used to further amplify the PCR product by an additional four cycles of PCR to introduce biotin tags to the 3′ ends of the amplified cDNAs. Approximately 300 ng cDNA was sonicated into 300-bp fragments using the Covaris S2 system and enriched with Dynabeads MyOneTM Streptavidin C1 beads. Libraries were constructed using a Kapa Hyper Prep Kit (Kapa Biosystems) and were then submitted to 150-bp paired-end sequencing on an Illumina Hiseq X Ten platform (Novogene).

For 10x Genomics single-cell RNA sequencing, we first used FACS to isolate living cells ($7AAD^-$ for CS11 and CS15 yolk sac), and then implemented the Chromium Single Cell 3′ v2 libraries, under the guidance of the official instruction manual (https://support.10xgenomics.com/single-cell-gene-expression/library-prep/doc/technical-note-assay-scheme-and-configuration-of-chromium-single-cell-3-v2-libraries).

Pre-processing scRNA-seq data

For scRNA-seq data, the raw reads from each cell were first split by the specific barcode sequence attached in Read 2. The template switch oligo (TSO) sequence and polyA tail sequence were trimmed for the corresponding Read 1 after UMI information was aligned to it. Subsequently, reads with adaptor contaminants or low-quality bases ($n > 10\%$) were discarded. Next, the stripped Read1 sequences were aligned to the hg19 human transcriptome (UCSC) using Hisat2 (version 2.1.0)[38]. Uniquely mapped reads were counted using the HTSeq package[39] and grouped by the cell-specific barcodes. Duplicated transcripts were removed based on the UMI information for each gene. Finally, for each individual cell, the copy number of transcripts of a given gene was the number of distinct UMIs for that gene.

To filter low quality cells, count values for each cell were first grouped into an expression matrix; only cells with more than 2,000 genes and 10,000 transcripts detected and below 50% transcripts mapped on the External RNA Controls Consortium (ERCC) reference were retained. Also, cells with too many raw reads (>1,000,000) and genes (>10,000) were excluded because these cells might not be real single cells. We obtained a total of 1,461 cells after sequencing, of which only 119 cells (8.15% of total) were excluded for failing the quality control threshold (Supplementary Table 1), leaving us with a median UMI of 132,783 and gene number of 4,860. The distribution of the UMI can be seen in Extended Data Fig. 1a, with the majority of the cells that showed more than 50% of UMIs mapping to the ERCCs falling below the UMI threshold of 1×10^4, while cells above 1×10^4 showed robust expression of housekeeping genes (Extended Data Fig. 1b).

Regarding the droplet-based scRNA-seq data from 10x Genomics, the dataset was aligned and quantified using the CellRanger software package (version 2.1.0) with default parameters, giving a total of 11,944 cells from the CS11 and CS15 yolk sacs. For 10x Genomics, we adopted a more relaxed quality control standard because the sequencing depth is lower than for STRT-seq. Cells with >1,000 genes expressed were retained, resulting in 25 cells excluded in total. The droplet-based scRNA-seq was expected to generate cell doublets at a low frequency, which could be incorrectly interpreted as novel cell types, and so to avoid the effects of doublets, 480 cells identified using Doubletdetection (https://doi.org/10.5281/zenodo.2678042) were removed from our data. We finally removed 1,874 cells with matured erythrocyte characteristics, leaving a total of 9,565 cells for our final analysis.

Cell-type detection and dimensionality reduction

Downstream analysis for well-based modified STRT-seq, such as data normalization, clustering, differential expression analysis, was implemented using the R package Seurat 2. After quality control, 1,342 cells were retained. To start with, preliminary UMAP and clustering analysis were performed, after which stromal cells (111 cells), including epithelial cells and mesenchymal cells, were identified on the basis of expression of *EPCAM* and *PDGFRA* and removed from our data. After quality control, 1,231 cells were retained. To start with, 888 highly variable genes (HVGs) were detected using the 'FindVariableGenes' function with the parameters of y.cutoff = 1 and x.low.cutoff = 1. These HVGs were used to perform PCA and the top 20 significant PCs were selected using the elbow of standard deviations of principal components (PCs). Next, selected PCs were used to perform UMAP analysis and cluster detection using the RunUMAP and FindClusters functions, respectively. Finally, eight progenitor clusters were annotated as YSMP, ErP, MkP, GMP, myeloblast, $CD7^{lo}P$, $CD7^{hi}P$ and HSPC, characterized by high expression of *CD34* and *MYB* and expressing specific genes for ErP (*GATA1*, *KLF1*), MkP (*GATA1*, *PF4*), GMP (*MPO*), myeloblast (*LYZ*), $CD7^{lo}P$ (*IL7R*, *CD7*), $CD7^{hi}P$ (*IL7R, CD7*) and HSPC (*HOXA6*, *HOXA10*). In addition, seven mature cell types were identified and defined as monocytes (*CCR2*, *HLA-DRA*), macrophages (*CD14*, *MRC1*), ILCs (*RORC*, *LTA*) and mast cells (*CPA3*, *CMA1*) for further downstream analysis.

To decode the cell type heterogeneity of the myeloid group, cells of unrelated types (ErP, MkP, mast cells, $CD7^{lo}P$, $CD7^{hi}P$, ILC and HSPC) were excluded from downstream analysis. We detected 2,195 HVGs with setting parameters x.low.cutoff = 0.3 and y.cutoff = 0.75, then selected the top 30 PCs to find clusters and to implement UMAP analysis, which resulted in 14 discrete cell clusters. Among the 14 clusters, in addition to YSMP, GMP, myeloblasts and monocytes, 10 macrophage clusters were identified and recognized as Blood_Mac, Liver_Mac, Skin_Mac, Lung_Mac, YS_Mac1, YS_Mac2, Head_Mac1, Head_Mac2, Head_Mac3, Head_Mac4, based on sampling site information.

For droplet-based 10x Genomics data, the pooled raw data from CS11 and CS15 yolk sacs were objected to Seurat 3 for data integration based on identification of 'anchors' between pairs of datasets following the tutorial at https://satijalab.org/seurat/v3.1/integration.html. In brief, SCTransform normalization was implemented separately for each dataset, after which the top 3,000 feature genes were selected and used for data integration. Next, the integrated data were used for dimensionality reduction and cluster detection. To start with, PCA was performed and the top 30 PCs were used for UMAP analysis using the umap package; then eight discrete clusters were detected using FindClusters with setting parameters dims = 1:10, resolution = 0.1 and annotated as YSMP (*CD34*, *MYB*), ErP/MkP (*GATA1*, *PF4*, *KLF1*), Mac (*CD14*, *CD163*), Endo (*CDH5*, *SOX7*), Epi (*EPCAM*) or Mes (*PDGFRA*) on the basis of feature genes. The same analysis procedures were applied for yolk sac dataset integration between modified STRT–seq and 10x data.

To identify sub-clusters in the Mac clusters, macrophage clusters were separately reanalysed. The subset data were normalized and the top 2,000 HVGs were recognized after ranking by residual variance using the 'vst' method through the SCTransform function. All HVGs were imported into PCA analysis carried out by RunPCA next.

SCENIC and GSEA analysis

To further assess the transcriptional and regulatory characteristics of the different progenitors including YSMP, ErP, MkP, myeloblast, $CD7^{lo}P$, $CD7^{hi}P$ and HSPC populations, gene regulatory network analysis was performed using SCENIC[40]. First, regulatory modules (regulons) were inferred from co-expression and DNA motif analysis. These regulons were then evaluated in each cell to ascertain their activity before a binary matrix was obtained. To profile the gene regulatory module features of all progenitors, the Spearman correlation coefficient between regulons was calculated, and only regulons with a correlation coefficient larger than 0.3 with at least one other regulon and activated in at least 30% cells in any progenitor clusters were included for visualization.

To find statistically significantly different regulons between YSMP and HSPC, gene set enrichment analysis (GSEA)[41] was performed on the regulon gene sets produced by SCENIC analysis, and the top two regulons ordered by *P* value were used for visualization for each population.

To compare the differential expression of erythroid and myeloid signature between YSMP in human and EMP in mouse, we performed GSEA analysis, in which the erythroid and myeloid gene sets refer to published data[42].

Developmental pseudotime analysis

The Monocle 2 package (version 2.8.0)[43,44] in R was used to determine the pseudotime of YSMP development in liver. First, cells sampled from liver (CS11 to CS17) belonging to the clusters related to myeloid lineage and excluding mature macrophages, namely YSMP, GMP, myeloblast, and monocyte, were selected and subjected to Monocle 2. Then, UMI data and HVGs obtained by FindVaribleGenes (x.low.cutoff = 0.5, y.cutoff = 0.5) in Seurat were input for unsupervised ordering of the cells. The remaining parameters were default. To find genes that changed their expression during the process of monocyte specification, we calculated Spearman correlations between predicted pseudotime

and gene expression: only genes with Spearman correlations higher than 0.4 were selected for visualization.

To interpret macrophage specification in head, only head macrophages (Head_Mac0, Head_Mac1, Head_Mac2, Head_Mac3 and Head_Mac4) were included and projected onto the UMAP plot. Clearly, head macrophages are ordered by sampling stage. To detect genes with dynamic changes in expression during head macrophage specification, we performed analysis of variance (ANOVA, one-sided). A total of 2,438 DEGs were identified and subjected to partitioning around medoids (PAM) analysis, a more robust K-means version, resulting in five gene expression patterns.

To predict the developmental trajectory of liver macrophages, gene expression data from all macrophages sampled from the liver were extracted. First, PCA was implemented after data were normalized and log transformed. Then, the top ten PCs were selected to create a diffusion map of cells through the destiny package with default parameters[45], after which DPT analysis was used to align cells in pseudotime order with setting cells of CS12 as the start point.

Signature gene detection

To find the signature genes of YSMPs, we compared YSMPs with other haematopoietic cells in the yolk sac and obtained DEGs with adjusted $P < 0.05$ and fold change >2.

We also probed signature genes in 10x data when comparing YSMPs (YSMP1 and YSMP2) with other haematopoietic clusters (ErP/MkP and Mac), before selecting DEGs with an adjusted $P < 0.05$ and fold change >1.5.

DEG and cluster biomarker identification

DEGs were identified by running the 'FindAllMarkers' function in Seurat. All DEGs of specific clusters are listed in the Supplementary Tables.

The datasets of surface markers and transcription factors were downloaded from Cell Surface Protein Atlas (http://wlab.ethz.ch/cspa/) and HumanTFDB3.0 (http://bioinfo.life.hust.edu.cn/HumanTFDB/) respectively.

Statistics and reproducibility

For all human embryos, specific sample information can be found in Fig. 1a and Supplementary Table 1.

For the STRT–seq analysis of human embryos, a total of $n = 8$ biologically independent samples across 7 time points were included. The time points and anatomical locations are CS11 (head and yolk sac, $n = 1$), CS12 (head, liver and yolk sac, $n = 1$), CS13 (head, blood and yolk sac, $n = 1$), CS15 (head, blood and liver, $n = 2$; yolk sac, $n = 1$), CS17 (head, blood, liver and yolk sac, $n = 1$), CS20 (brain, blood, skin, liver and lung, $n = 1$), and CS23 (brain, liver, skin, lung, $n = 1$). The total number of cells used in the final analysis was 1,231 (Figs. 1b–d, 2a–e, 3a–e, 4a–d and Extended Data Figs. 1–3, 5–8).

For the STRT–seq analysis of paediatric skin, a total of $n = 2$ biologically independent samples were obtained after elective circumcision. In total, 37 cells were used in the final analysis (Extended Data Fig. 8).

For the 3′ 10x data from yolk sac, a total of $n = 2$ biologically independent samples were included, at CS11 ($n = 1$) and CS15 ($n = 1$). In total, 9,565 cells were used in the final analysis (Extended Data Fig. 3).

Detailed cell numbers for scRNA-seq in each figure are shown below, with those from 10x data specifically indicated. Figure 1b, c and Supplementary Table 2: 1,231 haematopoietic cells (116 YSMP cells, 72 ErP cells, 30 MkP cells, 45 GMP cells, 100 myeloblast cells, 120 monocyte cells, 69 Mac_1 cells, 196 Mac_2 cells, 57 Mac_3 cells, 79 Mac_4 cells, 33 HSPCs, 104 $CD7^{hi}$P cells, 140 $CD7^{lo}$P cells, 21 ILCs, and 49 mast cells); Fig. 1d: 238 yolk sac haematopoietic cells and 354 liver haematopoietic cells; Fig. 2a–e and Supplementary Table 5: 88 liver haematopoietic cells; Fig. 3a: 782 myeloid cells (128 YSMP cells, 37 GMP cells, 103 myeloblast cells, 64 monocyte cells, 7 Liver_Mac cells, 51 Blood_Mac cells, 71 Lung_Mac cells, 46 Skin_Mac cells, 61 YS_Mac1 cells, 29 YS_Mac2 cells, 73 Head_Mac1 cells, 38 Head_Mac2 cells, 41 Head_Mac3 cells, and 33 Head_Mac4 cells); Fig. 3b, c: 450 macrophages (7 Liver_Mac cells, 51 Blood_Mac cells, 71 Lung_Mac cells, 46 Skin_Mac cells, 61 YS_Mac1 cells, 29 YS_Mac2 cells, 73 Head_Mac1 cells, 38 Head_Mac2 cells, 41 Head_Mac3 cells, and 33 Head_Mac4 cells); Fig. 3b: DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with P value adjusted for multiple testing using Bonferroni correction), genes with fold change >1.5 and adjusted $P < 0.05$ were selected, and the top three DEGs for each cluster were used for visualization; Fig. 3d, e: 514 myeloid cells with CS 20 and CS 23 cells excluded; Fig. 4a–d and Supplementary Table 7: 155 macrophages in head (9 Head_Mac0 cells, 45 Head_Mac1 cells, 29 Head_Mac2 cells, 39 Head_Mac3 cells and 33 Head_Mac4 cells).

Extended Data Figure 1a–c: 1,461 cells from $n = 8$ biologically independent embryo samples; Extended Data Fig. 1d–j: 1,231 haematopoietic cells; Extended Data Fig. 1j; DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with P value adjusted for multiple testing using Bonferroni correction), genes with fold change >1.5 and adjusted $P < 0.05$ were selected, and the top five DEGs for each cluster were used for visualization; Extended Data Fig. 1k, l: 188 progenitor cells (118 mouse EMP cells and 70 human YSMP cells); Extended Data Fig. 2a, d, f and Supplementary Table 3: 640 progenitor cells (116 YSMP cells, 72 ErP cells, 30 MkP cells, 45 GMP cells, 100 myeloblast cells, 33 HSPCs, 104 $CD7^{hi}$P cells, and 140 $CD7^{lo}$P cells); Extended Data Fig. 2b, c: 116 YSMP cells and 33 HSPC cells; Extended Data Fig. 2c, f: DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with P value adjusted for multiple testing using Bonferroni correction), genes with fold change >1.5 (1.25 for Extended Data Fig. 2c) and adjusted $P < 0.05$ were selected, and the top ten DEGs (top five for Extended Data Fig. 2f) for each cluster were used for visualization; Extended Data Fig. 2e: 259 progenitor cells in liver and 131 progenitor cells in blood; Extended Data Fig. 3a: 11,944 cells from 10x data; Extended Data Fig. 3b, c: 238 yolk sac cells from STRT–seq; Extended Data Fig. 3d, e and sheet 1 of Supplementary Table 4: 9,565 yolk sac cells from 10x data; Extended Data Fig. 3f: 93 macrophages from STRT–seq and 1,259 macrophages from 10x; Extended Data Fig. 3g, h: 9,803 cells (6,449 cells in CS11 yolk sac from 10x data, 3,116 cells in CS15 yolk sac from 10x data, and 238 yolk sac cells from STRT–seq); Extended Data Fig. 3h: DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with P value adjusted for multiple testing using Bonferroni correction), surface marker genes with fold change >1.25 and adjusted $P < 0.05$ were selected, and the top five DEGs for each cluster were used for visualization; Extended Data Fig. 5a–d: 88 liver haematopoietic cells from 4 biologically independent embryo samples; Extended Data Fig. 5e–g: 66 liver haematopoietic cells; Extended Data Fig. 6a–d and Supplementary Table 6: 782 myeloid cells (128 YSMP cells, 37 GMP cells, 103 myeloblast cells, 64 monocyte cells, 7 Liver_Mac cells, 51 Blood_Mac cells, 71 Lung_Mac cells, 46 Skin_Mac cells, 61 YS_Mac1 cells, 29 YS_Mac2 cells, 73 Head_Mac1 cells, 38 Head_Mac2 cells, 41 Head_Mac3 cells, and 33 Head_Mac4 cells); Extended Data Fig. 6c: DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with P value adjusted for multiple testing using Bonferroni correction), genes with fold change >1.5 and adjusted $P < 0.05$ were selected, and the top five DEGs for each cluster were used for visualization; Extended Data Fig. 6e: 450 macrophages (7 Liver_Mac cells, 51 Blood_Mac cells, 71 Lung_Mac cells, 46 Skin_Mac cells, 61 YS_Mac1 cells, 29 YS_Mac2 cells, 73 Head_Mac1 cells, 38 Head_Mac2 cells, 41 Head_Mac3 cells, and 33 Head_Mac4 cells); Extended Data Fig. 7a: 64 monocytes; Extended Data Fig. 7b: 176 monocytes and macrophages in head; Extended Data Fig. 7c: 64 monocytes and macrophages in lung; Extended Data Fig. 7d–g and Supplementary Table 9: 41 macrophages in liver; Extended Data Fig. 7h: 49 macrophages in skin; Extended Data Fig. 8a–c, e and Supplementary Table 8: 464 macrophages (20 adult head macrophages, 39 embryonic head macrophages, 97 adult liver macrophages, 9 embryonic liver macrophages, 160 adult lung macrophages, 59 embryonic lung macrophages, 37 paediatric skin macrophages, 43 embryonic skin

macrophages); Extended Data Fig. 8d: 464 macrophages (58 head macrophages, 104 liver macrophages, 156 lung macrophages, 45 skin macrophages, and 101 unspecified macrophages).

For Supplementary Tables 2–4, 6 and 8, DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with *P* value adjusted for multiple testing using Bonferroni correction). For Supplementary Tables 7 and 9, one-sided ANOVA was used to find DEGs between five head macrophage sub-clusters (Head_Mac0, Head_Mac1, Head_Mac2, Head_Mac3, and Head_Mac4) and four liver macrophage sub-clusters at different stages (CS12, CS15, CS17, and CS23), with *P* value adjusted using FDR using p.adjust function in R.

For cell culture of yolk sac populations, a total of $n = 4$ biologically independent embryo samples were used. For bulk culture of YSMP populations, 3 biologically independent samples (CS11, $n = 1$ and CS12, $n = 2$) were used, of which one CS12 sample was used for bulk co-culture of the $CD45^{+}CD34^{-}CD44^{-}$ population. For the single-cell functional assay of YSMPs, one CS13 yolk sac was used, and 39 single-cell wells with haematopoietic clusters were analysed by FACS to detect haematopoietic lineages generated (Fig. 1e, Extended Data Fig. 4a–e).

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Raw data from scRNA-seq analysis have been deposited in the NCBI Gene Expression Omnibus (GEO) under accession numbers GSE133345 and GSE137010. Source Data for four Figures and eight Extended Data Figures are provided within the online content of this paper.

Code availability

All data were analysed with standard programs and packages, as detailed above. Scripts can be found at https://github.com/yandgong307/human_macrophage_project.

34. Moore, K. L., Persaud, T. V. N. & Torchia, M. G. *The Developing Human E-Book: Clinically Oriented Embryology With STUDENT CONSULT Online Access* (Elsevier Health Sciences, 2011).
35. Li, L. et al. Single-cell RNA-seq analysis maps development of human germline cells and gonadal niche interactions. *Cell Stem Cell* **20**, 858–873.e854 (2017).
36. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
37. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protocols* **9**, 171–181 (2014).
38. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
39. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
40. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
41. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
42. Chambers, S. M. et al. Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell* **1**, 578–591 (2007).
43. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
44. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
45. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
46. Darmanis, S. et al. A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl Acad. Sci. USA* **112**, 7285–7290 (2015).
47. Cohen, M. et al. Lung single-cell signaling interaction map reveals basophil role in macrophage imprinting. *Cell* **175**, 1031–1044.e1018 (2018).
48. MacParland, S. A. et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* **9**, 4383 (2018).

Acknowledgements We thank F. Tang, Q. Li and Y. Hu from Peking University for technical assistance during single cell library construction, and L. Robinson for language editing of the manuscript. This study was supported by grants from the National Key Research and Development Program of China, Stem Cell and Translational Research (2017YFA0103401, 2016YFA0100601 and 2019YFA0110201), the National Natural Science Foundation of China (31425012, 31930054, 81890991, 31871173, 81800102, 81600077, 81900115 and 31800978), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2017ZT07S347), the Key Research and Development Program of Guangdong Province (2019B020234002), the Beijing Municipal Science & Technology Commission (Z171100000417009 and Z171100001117159), the National Key Research and Development Plan Young Scientists Program (2017YFA0106000), the State Key Laboratory of Proteomics (SKLP-K201502) and the China Postdoctoral Science Foundation (2018M643373). F.G. is supported by Singapore Immunology Network (SIgN) core funding. F.G. is a European Molecular Biology Organization (EMBO) YIP awardee and is supported by the Singapore National Research Foundation Senior Investigatorship (NRFI) NRF2016NRF-NRFI001-02. C.Z.W.L. is supported by an A*STAR Graduate Scholarship.

Author contributions B.L., F.G. and Y.L. designed the study. Z. Bian and H.S. performed sample preparation and FACS with help from Y.Z., Z. Bai and Y.N. T.H., Z. Bian, H.S., C.L. and J.H. performed scRNA-seq with help from J.Z. and X.L. Z. Bai performed cell culture with help from L.B. and Y.Z. L.B., C.M., R.Z. and L.C. collected and prepared the samples. Y.G. and T.H. performed bioinformatics analysis with help from Z.L., B.L., Y.L., Z. Bian, C.Z.W.L., J.K.Y.C. and L.G.N. Z. Bian, C.Z.W.L., T.H., Y.G., H.S., Z. Bai, Y.L., F.G. and B.L. wrote the manuscript, with contributions from all authors.

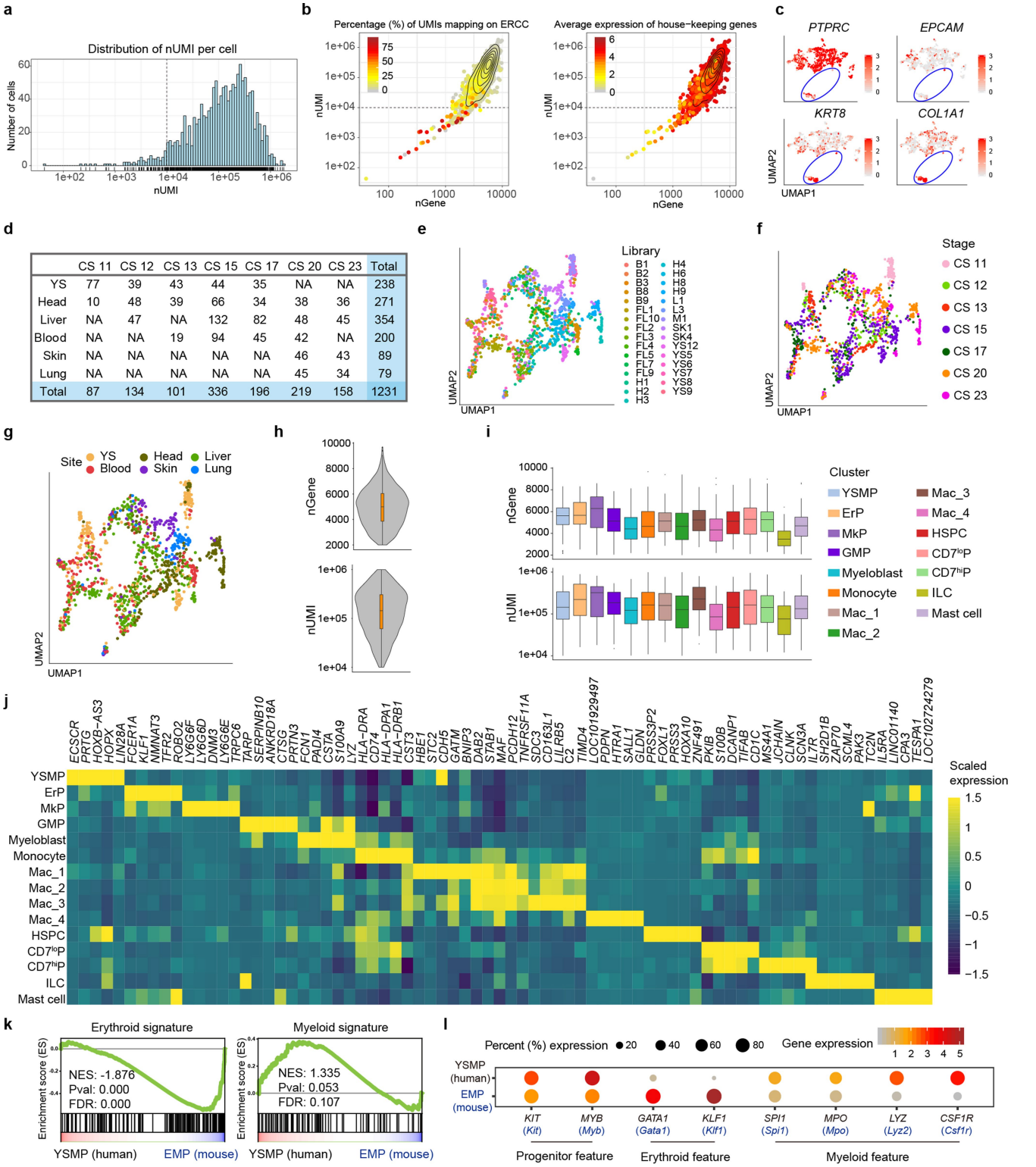
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2316-7.

Correspondence and requests for materials should be addressed to Y.L., F.G. or B.L.

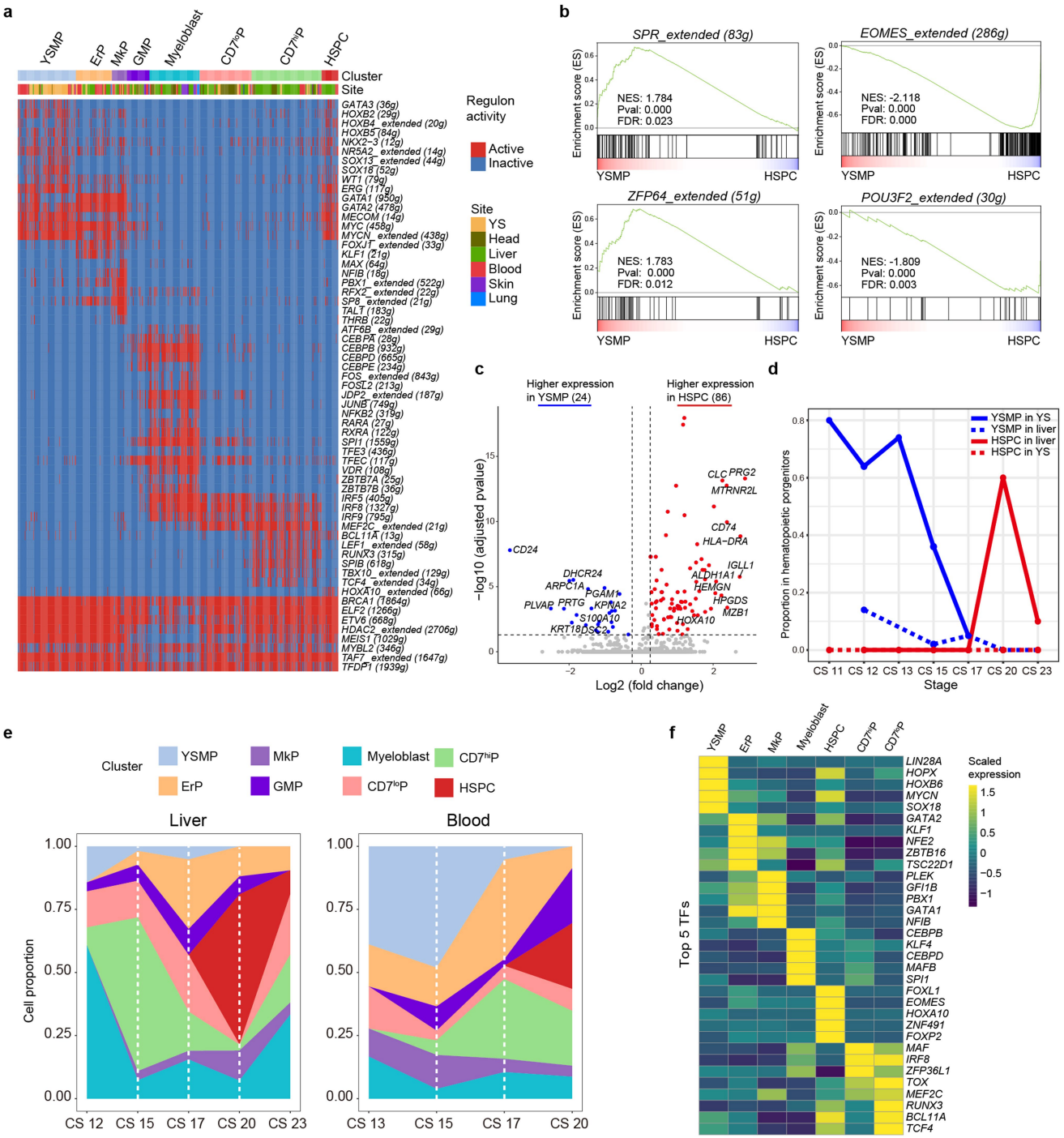
Peer review information *Nature* thanks Hans-Reimer Rodewald and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



	CS 11	CS 12	CS 13	CS 15	CS 17	CS 20	CS 23	Total
YS	77	39	43	44	35	NA	NA	238
Head	10	48	39	66	34	38	36	271
Liver	NA	47	NA	132	82	48	45	354
Blood	NA	NA	19	94	45	42	NA	200
Skin	NA	NA	NA	NA	NA	46	43	89
Lung	NA	NA	NA	NA	NA	45	34	79
Total	87	134	101	336	196	219	158	1231

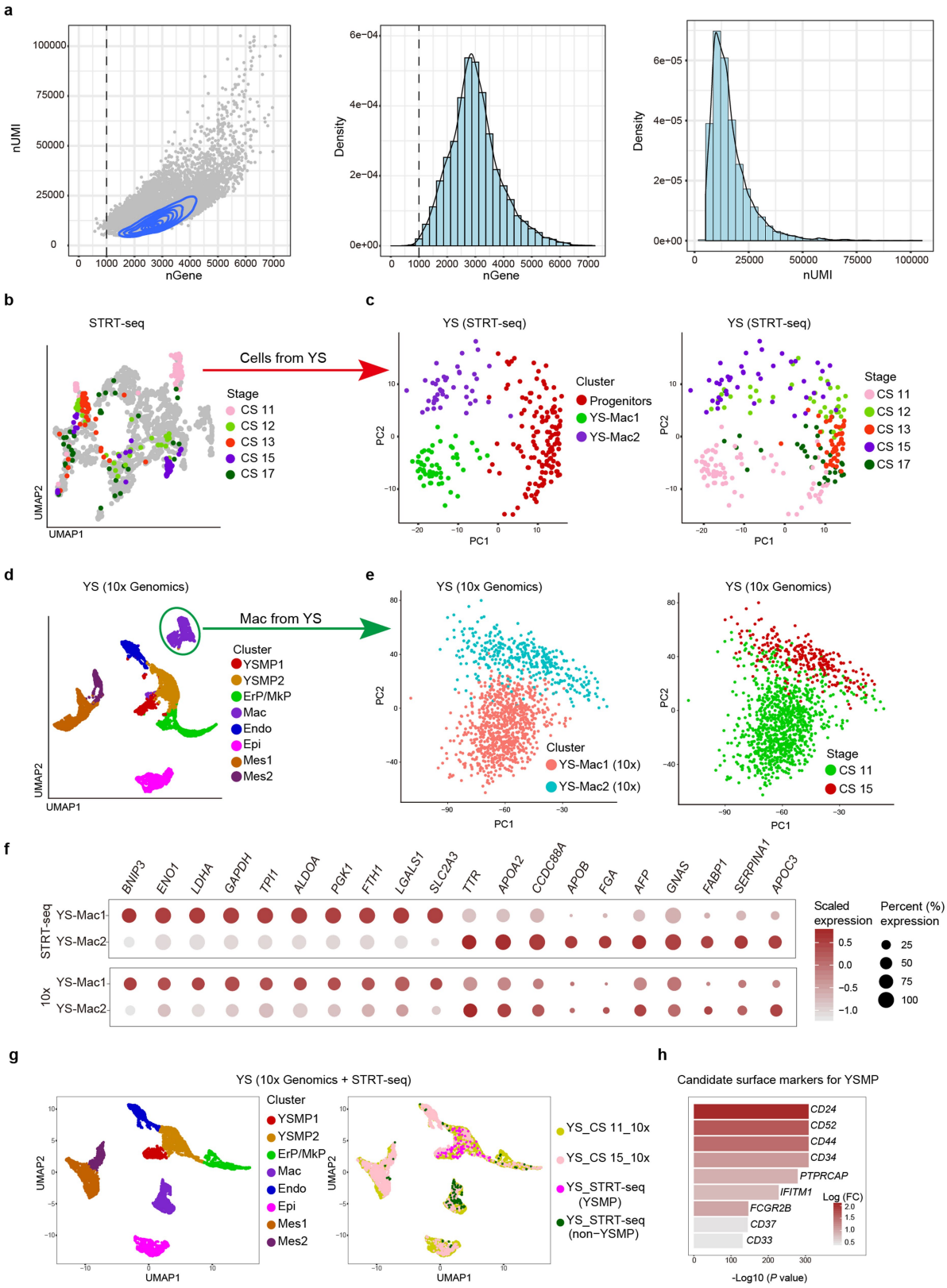
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Technical information about scRNA-seq library and overview of haematopoietic clusters. a, Graph showing distribution of UMI per cell for STRT–seq data ($n = 8$ biologically independent embryo samples and 1,461 cells). Threshold for final analysis was set at >10,000. **b**, Percentage of UMIs mapped on to ERCC and average expression of house-keeping genes. nGene: number of genes expressed, nUMI: number of UMIs expressed. **c**, UMAP visualization showing exclusion of non-haematopoietic cells from final analysis based on expression of *PTPRC*, *EPCAM*, *KRT8* and *COL1A1*. **d**, Numbers, location and Carnegie stage information for cells used in final analysis. **e**, UMAP showing minimal batch effect among single cell libraries ($n = 1,231$ cells). **f**, **g**, UMAP visualization of all haematopoietic cells with Carnegie stage (**f**) and site (**g**) information mapped on. **h**, Violin plots of average gene and UMI numbers of scRNA-seq in STRT–seq data. For box plot within each violin plot, centre black lines indicate median values, boxes range from 25th to 75th percentiles, and whiskers correspond to 1.5× IQR. **i**, Gene and UMI numbers of identified haematopoietic clusters. Centre black lines indicate median values, boxes range from 25th to 75th percentiles, and whiskers correspond to 1.5× IQR. **j**, Heat map of the top five DEGs between haematopoietic clusters. DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with *P* value adjusted for multiple testing using Bonferroni correction), and genes with fold change >1.5 and adjusted $P < 0.05$ were selected. **k**, GSEA plots of erythroid and myeloid signatures indicate significantly lower expression of erythroid signature in human YSMPs than in mouse EMPs (extracted from a published scRNA-seq dataset)[19] ($n = 70$ human YSMP cells and 118 mouse EMP cells). *P* value was calculated using permutation test (one-sided) based on phenotype by GSEA 3.0 software, representing statistical significance of normalized enrichment score (NES). FDR, false discovery rate. **l**, Gene expression of haematopoietic progenitor feature (*CD34*, *MYB*), erythroid feature (*GATA1*, *KLF1*) and myeloid feature (*SPI1*, *MPO*, *LYZ*, *CSF1R*) in human YSMPs and mouse EMPs[13,19].



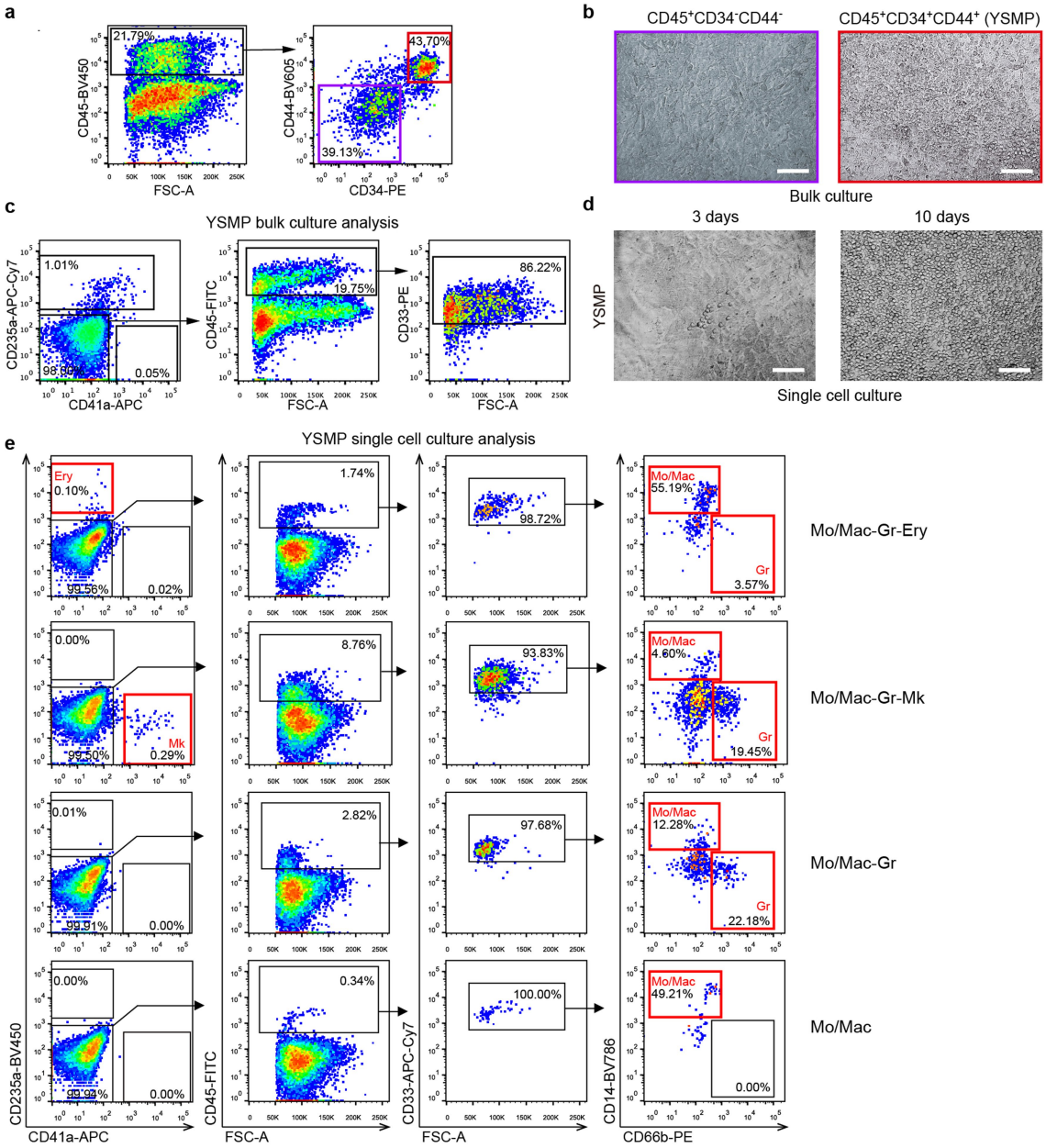
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Characteristics of haematopoietic progenitors in human embryos. a, Heat map showing differential regulon expression between haematopoietic progenitor clusters (YSMP, $n = 116$; ErP, $n = 72$; MkP, $n = 30$; GMP, $n = 45$; myeloblast, $n = 100$; $CD7^{lo}$ progenitor ($CD7^{lo}$P), $n = 140$; $CD7^{hi}$ progenitor ($CD7^{hi}$P), $n = 104$; HSPC, $n = 33$) generated by SCENIC and clear sets of cell-type specific regulons that may play critical roles in the development of each progenitor population. The number of genes associated with each regulon is listed in parentheses. ErP and MkP signatures were very similar, although MkP appeared to have downregulated expression of *KLF1*and up-regulated expression of other platelet-related transcription factors such as *TAL1* and *NFIB*. $CD7^{lo}$P had overlapping modules with myeloblast and GMP, sharing the myeloid-restricted *TFEC* pathways, but lacked expression of more myeloid-committed *CEBP*s. $CD7^{hi}$P showed many signatures typical of lymphoid potential, such as the activation of *LEF1* and *TCF4* signals. HSPC were characterized by activation of the *HOXA10* module, as well as higher levels of lymphoid-associated *BCL11A* when compared to YSMP. **b**, GSEA plots of the top two differentially expressed regulons between YSMP ($n = 116$ cells) and HSPC ($n = 33$ cells). GSEA analysis revealed that YSMPs highly enriched the *SPR* and *ZFP64* regulons, while HSPCs had higher expression of the *EOMES* and *POU3F2* modules. *P* value was calculated using permutation test (one-sided) based on phenotype by GSEA 3.0 software, representing the statistical significance of enrichment score. **c**, Volcano plot of DEGs between YSMP ($n = 116$ cells) and HSPC ($n = 33$ cells), with the top 10 genes for each cluster indicated. Although the regulon landscape was similar between these two groups, we identified 110 DEGs (Supplementary Table 3). There were more upregulated genes in HSPC (86, red) than in YSMP (24, blue), with HSPC expressing genes related to antigen presentation including *CD74* and *HLA-DRA* as well as lymphoid-related genes including *IGLL1*. DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with *P* value adjusted for multiple testing using Bonferroni correction), and genes with fold change >1.25 and adjusted $P < 0.05$ were selected. **d**, Proportion changes of YSMPs and HSPCs in the haematopoietic progenitor populations of yolk sac and liver between CS11 and CS23 ($n = 8$ biologically independent embryo samples). The proportion of the YSMP population peaked at CS11 before steadily decreasing, while that of the HSPC population expanded between CS17 and CS20 before reducing to about 10% at CS23. **e**, Proportion changes of different haematopoietic progenitor clusters from CS12 to CS23 in the liver ($n = 6$ biologically independent embryo samples and 259 cells), and CS13 to CS20 in the blood ($n = 5$ biologically independent embryo samples and 131 cells). **f**, Heat map showing expression levels of the top five differentially expressed transcription factors between YSMP, ErP, MkP, myeloblast, HSPC, $CD7^{lo}$P and $CD7^{hi}$P cells. DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with *P* value adjusted for multiple testing using Bonferroni correction), and genes with fold change >1.5 and adjusted $P < 0.05$ were selected.

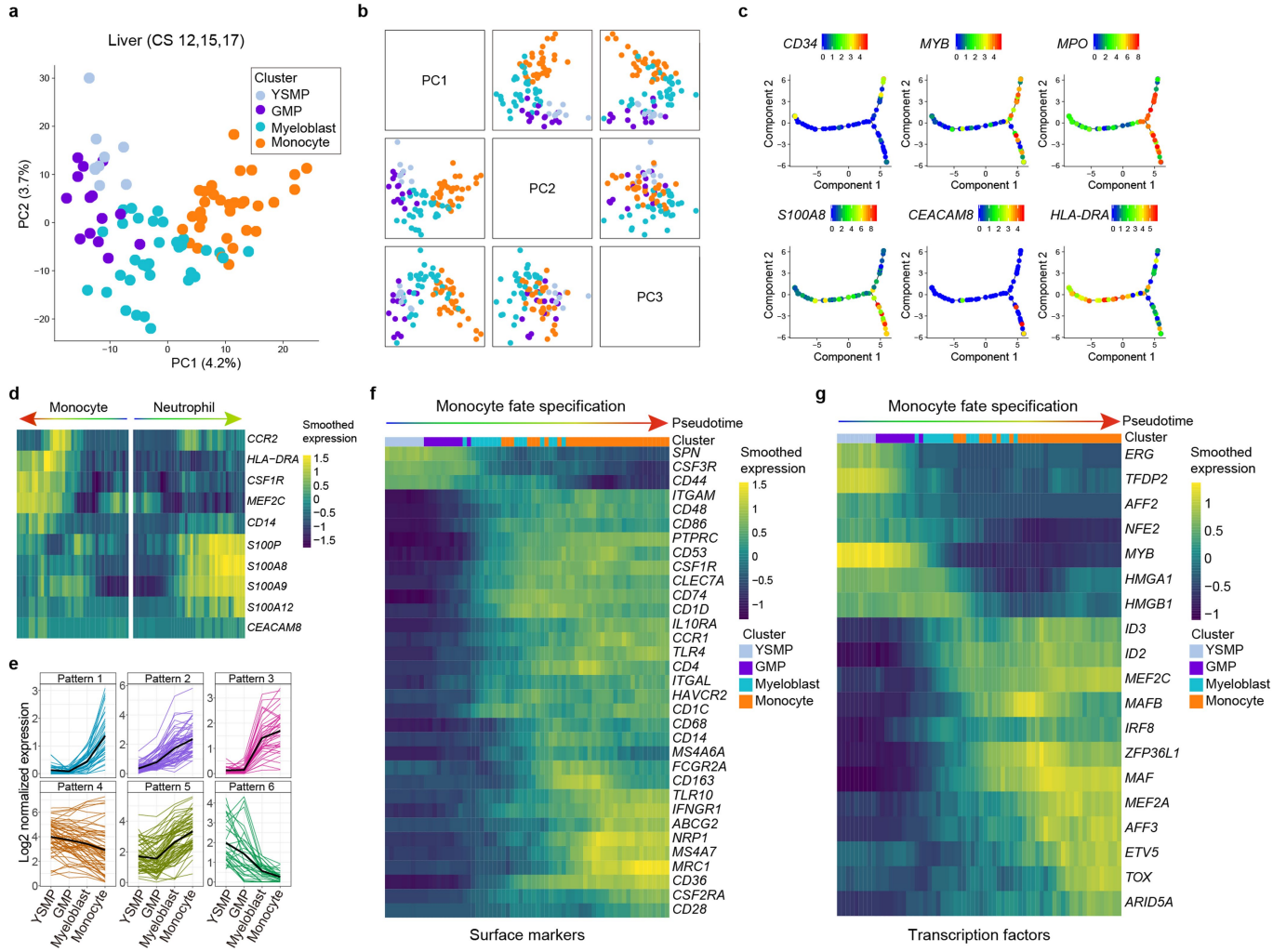


Extended Data Fig. 3 | See next page for caption.

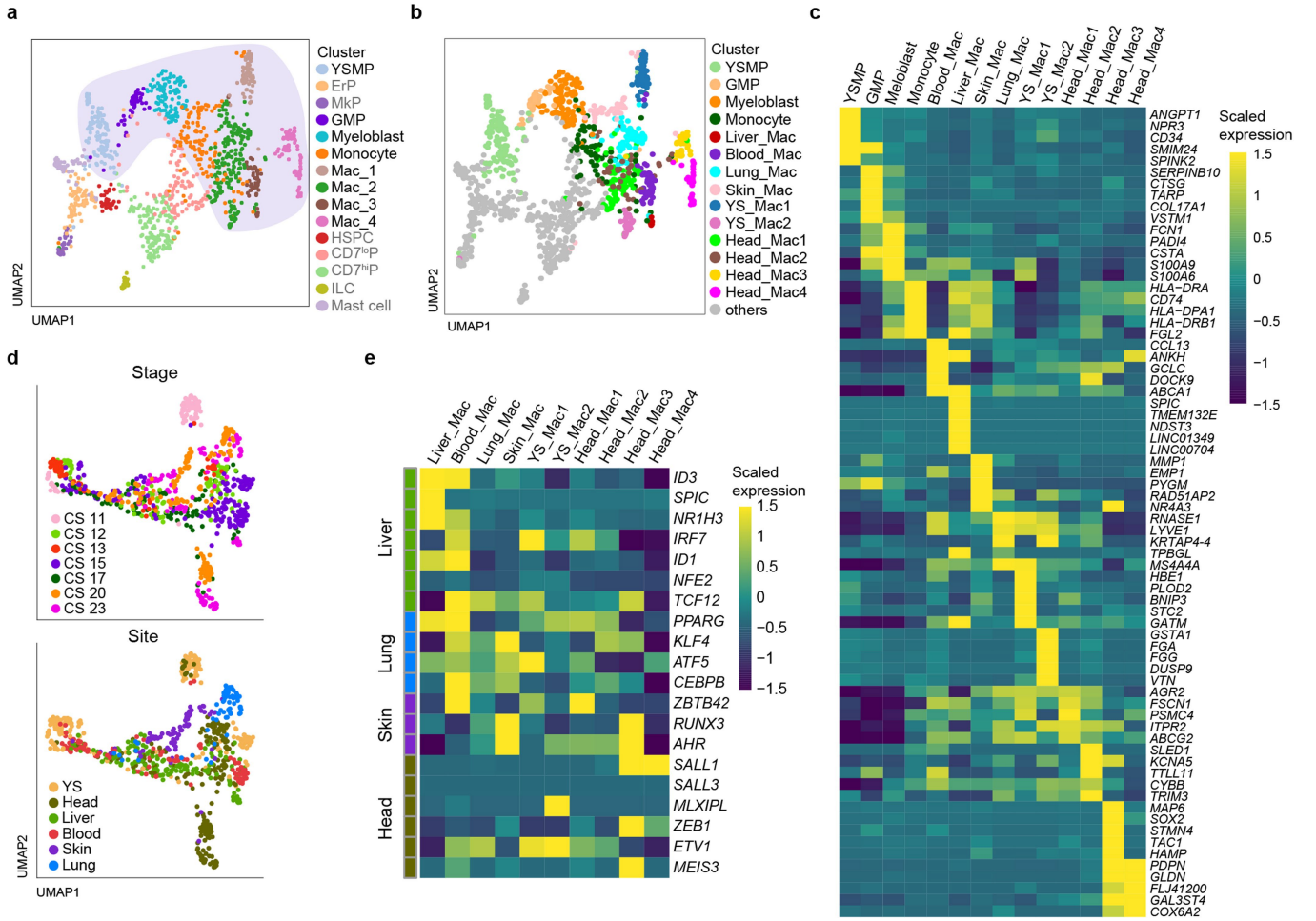
Extended Data Fig. 3 | Validation of STRT–seq macrophage clustering in yolk sac by 10x Genomics. a, Quality control for 10x Genomics data by UMI and gene numbers ($n = 2$ biologically independent embryo samples and 11,944 cells). The threshold for final analysis was set as gene number >1,000 per cell. **b**, UMAP visualization of total haematopoietic clusters generated via STRT–seq with cells from yolk sac ($n = 5$ biologically independent embryo samples and 238 cells) mapped on and coloured by stage information, which were extracted for further analysis. **c**, PCA of cells from yolk sac in STRT–seq with re-clustering (left) and stage (right) information mapped on ($n = 5$ biologically independent embryo samples and 238 cells). These cells were re-clustered into three clusters and annotated by gene expression profiles. The YS-Mac1 cluster mainly consisted of cells from CS11, while the YS-Mac2 cluster mainly consisted of cells from CS15. Based on these findings, we selected the 10x Genomics data from the CS11 and CS15 yolk sacs to validate our clustering. **d**, UMAP visualization of 10x Genomics data from CS11 and CS15 yolk sacs ($n = 2$ biologically independent embryo samples and 9,565 cells). The Mac cluster was extracted for further analysis. **e**, PCA of Mac cluster in 10x Genomics data with re-clustering (left) and stage (right) information mapped on. **f**, Expression profile of top ten DEGs between YS-Mac1 and YS-Mac2 identified by STRT–seq ($n = 238$ cells) and projected onto the 10x Genomics data ($n = 9,565$ cells). Even though 10x Genomics data have lower depth compared to STRT–seq, similar expression profiles can be seen for the majority of genes. **g**, UMAP visualization of integrated yolk sac data from STRT–seq and 10x Genomics analysis ($n = 9,803$ cells). There is less overlap in the mesenchymal (Mes1 and Mes2) and epithelial (Epi1) clusters because the STRT–seq data were only from $CD45^+$ haematopoietic cells, whereas the 10x Genomics data were from all yolk sac cells. Note that YSMPs from both datasets are well merged. **h**, Bar plot showing putative surface markers of YSMPs. On the basis of these data, CD34 and CD44 were selected for functional assays. DEGs were identified from combined STRT–seq and 10x Genomics data using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test with P value adjusted for multiple testing using Bonferroni correction), and surface marker genes with fold change >1.25 and adjusted $P < 0.05$ were selected.



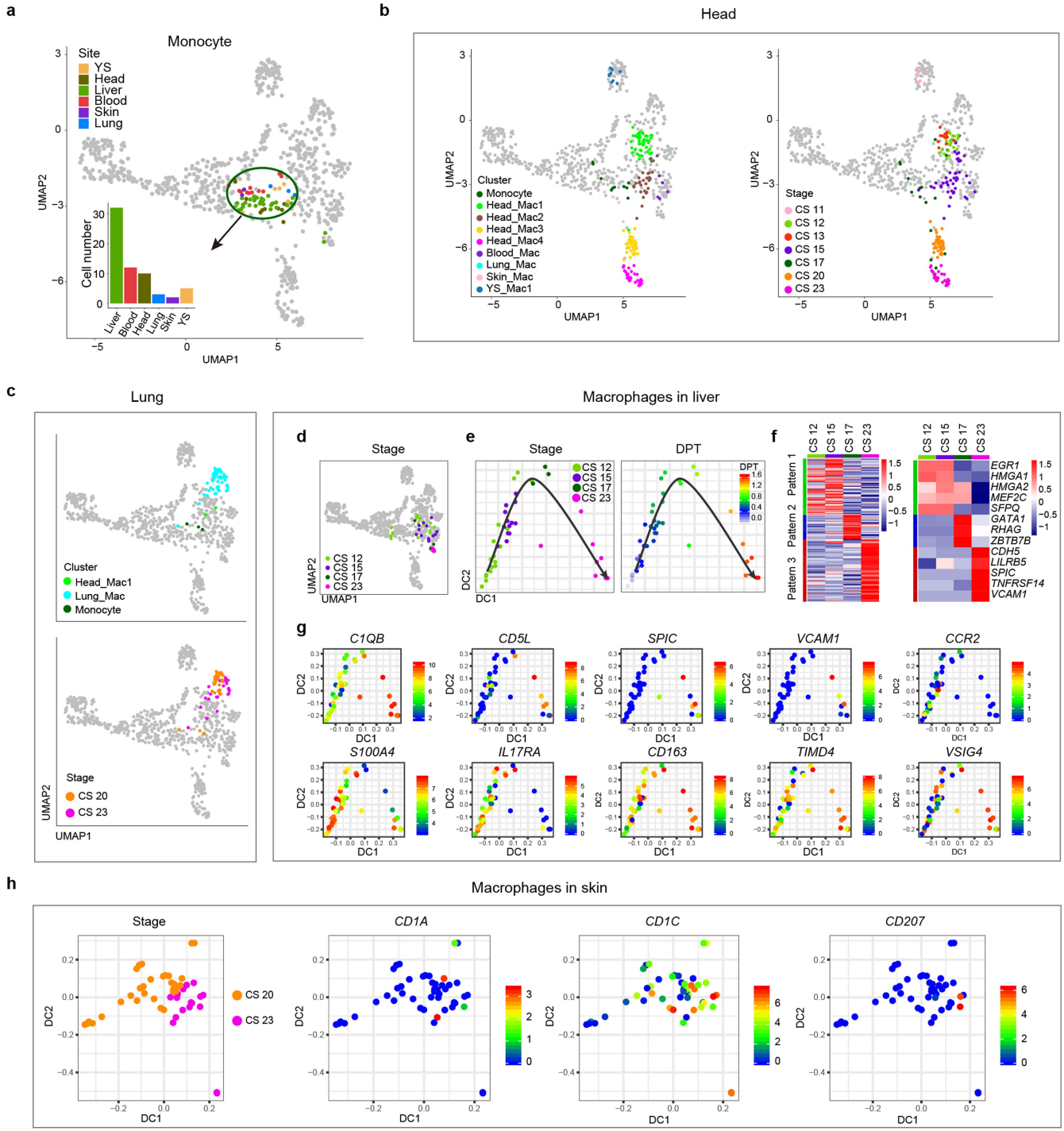
Extended Data Fig. 4 | In vitro functional assay of YSMPs. a, Gating strategy for sorting of the YSMPs ($CD45^+CD34^+CD44^+$) from a CS12 yolk sac. **b**, Representative morphologies of bulk cultures (100 cells per well) of negative control cells ($CD45^+CD34^-CD44^-$, $n = 5$ replication wells) and YSMPs ($CD45^+CD34^+CD44^+$, $n = 16$ replication wells) after 14 days of culture on MS5 feeder layer. $n = 3$ independent experiments from one sample of CS11 yolk sac and two samples of CS 12 yolk sac for YSMPs. Scale bars, 100 μm. **c**, Representative FACS analysis of cells collected from bulk cultures of YSMPs. Note that the myeloid cells ($CD33^+$) are predominant, in contrast to a small number of erythroid cells ($CD235a^+$) detected ($n = 3$ independent experiments). **d**, Representative morphologies of haematopoietic cells generated by a single YSMP from a CS13 yolk sac after 3 and 10 days of culture on MS5 feeder layer. In total, 184 YSMPs were individually cultured and 67 of them generated morphologically typical haematopoietic clusters. Scale bars, 100 μm. **e**, Representative FACS analysis of four kinds of distinct differentiation potential of single YSMPs. Cells were collected from the single-cell YSMP cultures and in total 39 wells were individually analysed. Lineage differentiation potentials are indicated in red for each clone. Mo/Mac, monocytes/macrophages ($CD45^+CD33^+CD14^+$); Gr, granulocytes ($CD45^+CD33^+CD66b^+$); Ery, erythrocytes ($CD235a^+$); Mk, megakaryocytes ($CD41a^+$).



Extended Data Fig. 5 | Developmental trajectory of YSMP in human embryonic liver. a, PCA of YSMP, GMP, myeloblast and monocyte populations sampled from CS12 to CS17 livers using PC1 and PC2 ($n = 4$ biologically independent embryo samples and 88 cells). **b**, PCA matrix of YSMP, GMP, myeloblast and monocyte populations. **c**, Monocle visualization of YSMP, GMP, myeloblast and monocyte populations sampled from CS12 to CS17 livers with the expression of the indicated genes mapped on. **d**, Heat map showing scaled expression of branching curated genes of monocyte and neutrophil fates ordered by pseudotime. **e**, Six main patterns of gene expression compared between YSMP, GMP, myeloblast and monocyte clusters. The expression levels of all pattern genes (coloured lines) and the average expression of each pattern (black line) are shown. The complete list of genes can be found in Supplementary Table 5. **f**, **g**, Heat maps displaying expression of monocyte specification-related surface markers (**f**) and transcription factors (**g**) along pseudotime.

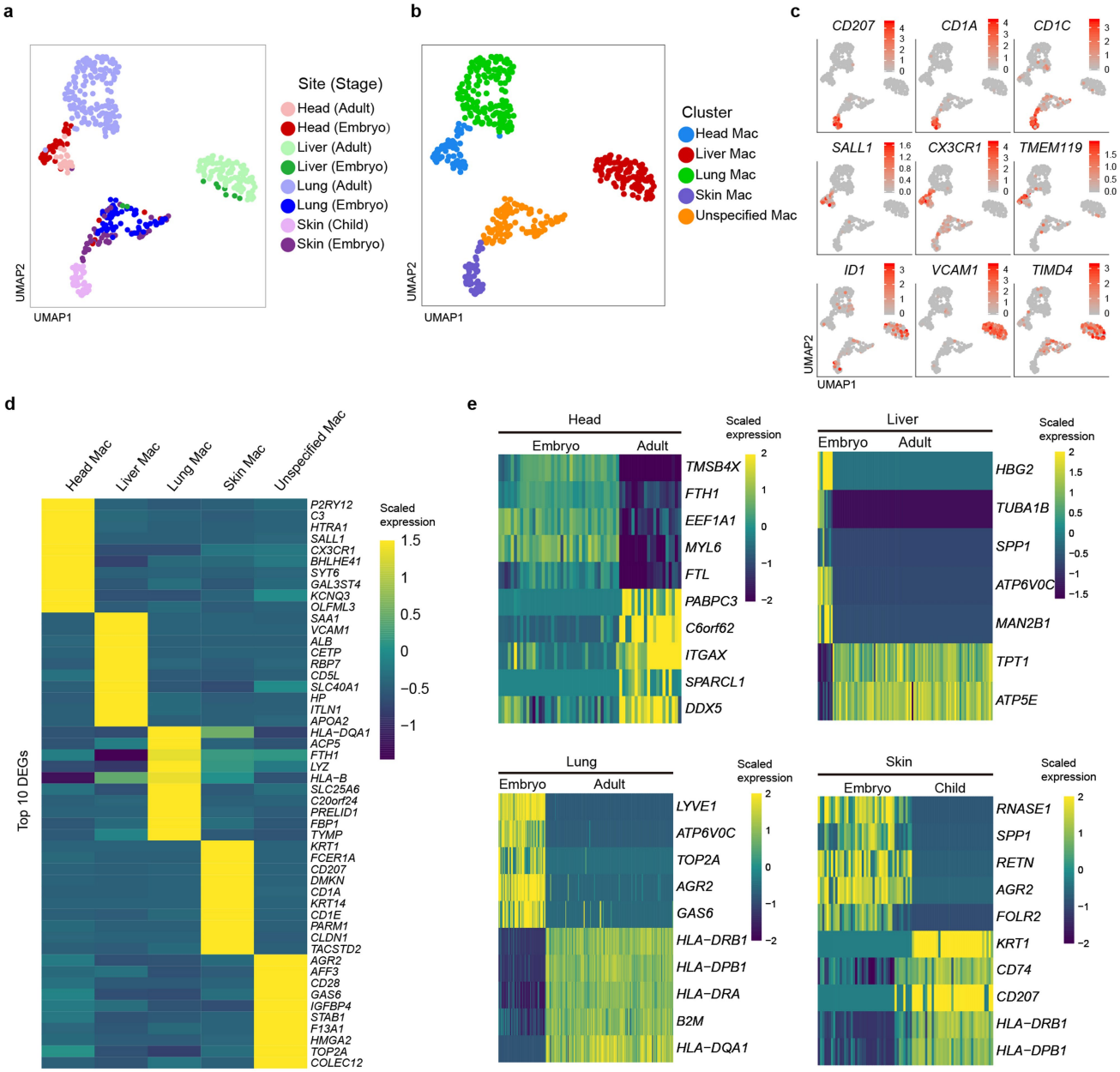


Extended Data Fig. 6 | Two distinct waves of yolk sac-derived macrophages contribute to TRM populations in human embryos. a, The area coloured lavender on the UMAP visualization highlights the myeloid groups selected for re-analysis ($n = 8$ biologically independent embryo samples and 1,231 cells). **b**, UMAP visualization of all the haematopoietic cells with re-clustered myeloid and macrophage clusters ($n = 782$ myeloid cells) mapped on. **c**, Heat map showing scaled expression of the top five DEGs for each re-clustered myeloid and macrophage population. DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with *P* value adjusted for multiple testing using Bonferroni correction), and genes with fold change >1.5 and adjusted $P < 0.05$ were selected. **d**, UMAP visualization of myeloid cells with stage (top) and site (bottom) information mapped on. **e**, Heat map showing scaled expression of curated TRM signature genes from a previous mouse study[24] in the re-clustered macrophage populations ($n = 450$ cells: 7 Liver_Mac cells, 51 Blood_Mac cells, 71 Lung_Mac cells, 46 Skin_Mac cells, 61 YS_Mac1 cells, 29 YS_Mac2 cells, 73 Head_Mac1 cells, 38 Head_Mac2 cells, 41 Head_Mac3 cells, and 33 Head_Mac4 cells).



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Monocyte distribution and macrophage specification in human embryonic head, lung, liver and skin. a, UMAP visualization of myeloid cells with monocytes ($n = 64$ cells) coloured by site information mapped on. Bar plot shows cell numbers at different sites. **b**, UMAP visualization of myeloid cells with monocytes and macrophages from human embryonic head ($n = 176$ cells) mapped on. Cluster (left) and stage information (right) are indicated by colours. **c**, UMAP visualization of myeloid cells with monocytes and macrophages from human embryonic lung ($n = 64$ cells) mapped on. Cluster (top) and stage information (bottom) is indicated by colours. **d**, UMAP visualization of the myeloid cells with macrophages from human embryonic liver ($n = 41$ cells) coloured by stage information mapped on. These cells were used to study Kupffer cell specification in situ. **e**, DiffusionMap visualizing differentiation trajectory of embryonic Kupffer cells with stage information (left) and pseudo-order (right) mapped on. Note that the cells also lined up in a continuum from CS12 to CS23, suggesting the gradual and sequential acquisition of TRM identity. **f**, Heat maps showing scaled expression of DEGs (left) and transcription factors within DEGs (right) in embryonic Kupffer cells across stages with three main gene expression patterns identified. DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with P value adjusted for multiple testing using Bonferroni correction), and genes with fold change >1.5 and adjusted $P < 0.05$ were selected. Complete gene list can be found in Supplementary Table 9. **g**, DiffusionMap visualizing differentiation trajectory of embryonic Kupffer cells with expression levels of the indicated genes mapped on. Expression of *C1QB*, a gene associated with macrophage tissue residency, was gradually upregulated, while genes related to Kupffer cell function such as *CD5L*, *SPIC* and *VCAM1* were expressed only at the end of the developmental pathway, suggesting that specialized Kupffer cells began to appear after CS17. Many of the downregulated genes are inflammation- or migration-related, such as *CCR2*, *S100A4* and *IL17RA*, while the expression of residency and Kupffer cell identity genes such as *CD163*, *TIMD4* and *VSIG4* was increased. Many of the signature genes, such as *SPIC* and *VCAM1*, have been previously reported in TRMs using animal models, which confirms that these cells were moving towards a more differentiated tissue-resident state. **h**, DiffusionMap visualization of macrophages from human embryonic skin ($n = 49$ cells) with stage information (left) and the expression levels of the indicated genes (right) mapped on.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Characteristics of human embryonic TRMs versus conventional TRMs. a, **b**, UMAP visualization of embryonic TRMs and their conventional TRM counterparts with site and stage (**a**) and cluster (**b**) information mapped on (n = 464 cells: 20 adult head macrophages, 39 embryonic head macrophages, 97 adult liver macrophages, 9 embryonic liver macrophages, 160 adult lung macrophages, 59 embryonic lung macrophages, 37 paediatric skin macrophages, and 43 embryonic skin macrophages). We performed combined analysis including the four embryonic TRM populations in the present study (Head_Mac4, Liver_Mac, Lung_Mac, and Skin_Mac) and the corresponding conventional TRMs in adults (head, liver and lung from public scRNA-seq data)[46–48] and children (skin). In total, five main macrophage clusters (head, liver, lung, skin and unspecified) were identified by unsupervised clustering. Two of the embryonic TRM populations (head and liver) clustered with their corresponding adult counterparts. The embryonic skin TRMs distributed into both specified and unspecified clusters, with the former cluster together with those from paediatric skin. The embryonic lung macrophages did not cluster at all with those in the adult lung, which indicated that the differentiation and specification of these TRMs had not yet occurred. **c**, UMAP visualizations of all TRM clusters with the expression levels of Langerhans (*CD207*, *CD1A* and *CD1C*), microglial (*SALL1*, *CX3CR1* and *TMEM119*) and Kupffer (*ID1*, *VCAM1* and *TIMD4*) cell-related genes mapped on. **d**, Heat map showing scaled expression of the top ten DEGs between the five identified human macrophage clusters (n = 464 cells: 58 head macrophages, 104 liver macrophages, 156 lung macrophages, 45 skin macrophages, and 101 unspecified macrophages). DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with P value adjusted for multiple testing using Bonferroni correction), and genes with fold change >1.5 and adjusted $P < 0.05$ were selected. Note that they were distinguished by the expression of TRM genes that have been well described in previous animal and human studies, such as *CD207* for the skin, *VCAM1* for the liver and *P2RY12* for the head. **e**, Heat map showing scaled expression of DEGs between the embryonic TRMs and their conventional TRM counterparts in each tissue (head, liver, lung and skin). DEGs were detected using FindAllMarkers function in Seurat (one-sided Wilcoxon rank-sum test, with P value adjusted for multiple testing using Bonferroni correction), and genes with fold change >1.5 and adjusted $P < 0.05$ were selected. The complete list of genes can be found in Supplementary Table 8. Many of the upregulated genes in the embryonic TRMs are related to cell cycle or tissue development, whereas the upregulated genes in the conventional TRMs are more related to immune function. For example, in the head, the embryonic TRMs expressed the neurodevelopmental gene *TMSB4X* as well as the cell cycle-related gene *EEF1A1*, whereas conventional TRMs expressed the immune-related gene *ITGAX*. Embryonic skin macrophages expressed the chemokine *SPP1*, indicating that they are either cells in transition or have newly arrived in the niche, further supporting our prediction that skin TRM specification has just begun at this time-point.

nature research

Corresponding author(s): Bing Liu, Florent Ginhoux, Yu Lan

Last updated by author(s): Mar 1, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	RS image expression (version 4.5.1.3)
Data analysis	Software used include: BD FACSDIVA (v8.01), Flowjo (v10), Cell Ranger (v2.10), Hisat2 (v2.10), HTSeq (v0.11.2), R (v3.50), Rstudio (v1.1), Seurat (v2.31 & v3.10), SCENIC (v0.96), Monocle 2 (v2.80), destiny (v2.10.0), GSEA (v3.0), Adobe Illustrator CS6 (16.0.0). Detailed parameters of each of the methods are mentioned in relevant sections in Methods. All data were analyzed with standard programs and packages, as detailed above. Scripts can be available at https://github.com/yandgong307/human_macrophage_project.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw data from scRNA-seq analysis have been deposited in the NCBI Gene Expression Omnibus under accession number GSE133345 and GSE137010. Source data for 4 figures and 8 extended data figures are provided within the online content of this paper.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size of scRNA-seq were determined by availability of human tissues. Final dataset scale was determined according to the quality control criteria as described in the methods.
Data exclusions	Quality control standard is established in advance according to the previous literature, which was accord with the general standard of the single cell library construction. We have used the following cell exclusion criteria: low quality, doublets, non-CD45+CD235a- cells, and the details placed in the section of methods.
Replication	In this study, we have 2 biologically independent CS 15 embryos and 2 biologically independent foreskins for STRT-seq. There were no other replications for STRT-seq and 10x. For cell culture of yolk sac populations, a total of n = 4 biologically independent embryo samples were used. For the bulk culture of YSMP population, n = 3 (one CS 11 and two CS 12) . For the single cell functional assay of YSMPs, n = 1 CS 13 yolk sac was used and n = 39 single-cell wells were analyzed by FACS to detect hematopoietic lineages generated. We used 10x Genomics and colony assay to verify the results found by STRT-seq, and found the cluster identity were consistent across different samples and sequencing techniques.
Randomization	We did not randomise the samples, as we were studying the trajectory of healthy embryonic hematopoietic development in general.
Blinding	Blinding is not relevant, as we are not studying pathology or disease.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

7-amino-actinomycin D (7-AAD) (PerCP-Cy5.5, 00699350 eBioscience, Lot 1910559)
anti-CD1a (APC, 559775 Biolegend, Lot 8164562)
anti-CD14 (BV786, 563698 BD, Lot 8351911)
anti-CD207 (PE, 564727 BD, Lot 8135683)
anti-CD235a (APC-Cy7, 349116 Biolegend, Lot 7355682 and B289027)
anti-CD235a (Pacific Blue, 306611 BioLegend, Lot B224563)
anti-CD33 (PE, 555450 BD, Lot 8074660)
anti-CD33 (APC-Cy7, 366614 BioLegend, Lot B252646)
anti-CD34 (PE, 550761 BD, Lot 7129824)
anti-CD41 (APC, 17-0419-42 eBioscience, Lot 2073742)
anti-CD44 (BV605, 562991 BD, Lot 7103609)
anti-CD45 (BV421, 563879 BD, Lot 9066960)
anti-CD45 (FITC, 11-0459-42 eBioscience, Lot 4310016)
anti-CD66b (PE, 561650 BD, Lot 7264511)

Validation	Antibodies used in the study are all commercial and verified by the manufacturer with catalog number and lot number provided.

Human research participants

Policy information about studies involving human research participants

Population characteristics	All human embryos included were between Carnegie stage 11 and 23. We collected 3 male, 6 female and 2 gender non-available embryos. The integrity of the samples was confirmed and examined at multiple steps. From the beginning, the Obstetrics and Gynecology clinicians confirmed the collected embryos and fetuses free of any known genetic or developmental abnormality. Before sample processing, the morphological examinations were performed to exclude samples with any potential development defects. The pediatric skin samples were obtained from foreskin of two children (8 and 10 years old).
Recruitment	The embryos were obtained from pregnant women undergoing drug abortion at The Fifth Medical Center of the PLA General Hospital (Beijing, China) after obtaining informed consent. Inclusion criteria: (1) Females older than 18 years of age and pregnant for 20-80 days (about 3-11 weeks) (2) Healthy pregnant women who intend to voluntarily terminate pregnancy through drug abortion and have no history of chronic diseases, genetic diseases and infectious diseases. Based on the principle of voluntary participation, no compensation had been paid in this study. Given the samples were extremely rare and precious, the sample selection was entirely determined by the availability of clinical specimens. Enrolled embryos were all quality controlled, which had been performed by two persons respectively. There was no self-selection bias in the progress of sample enrolled. Pediatric skin samples were obtained from tissue discarded after elective circumcision at Beijing Children's Hospital, Capital Medical University with parental consent.
Ethics oversight	All protocols were approved by the institutional review boards (The Affiliated Hospital of Academy of Military Medical Sciences Ethics Committee and Beijing Children's Hospital, Capital Medical University Ethics Committee; approval number: ky-2017-3-5 & IEC-C-008-A08-2018-75) and in accordance with the regulations of the Declaration of Helsinki. All the protocols were compliant with the Interim Measures for the Administration of Human Genetic Resources, administered by the Ministry of Science and Technology of China.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	The embryos were obtained from pregnant women undergoing drug abortion at The Fifth Medical Center of the PLA General Hospital (Beijing, China) after obtaining informed consent. The integrity and morphology of the embryos were evaluated, the somite pairs were counted, and the crown-rump length of the embryos were measured under the microscope to define the developmental stage1. The child skin was obtained from tissue discarded after elective circumcision at Beijing Children's Hospital, Capital Medical University with parental consent. Samples were immediately transported in RPMI1640 medium (Gibco,11875093) containing 10% Fetal Bovine Serum (HyClone, SH30070.03) on ice and processed within two hours. The various tissues (yolk sac, head, lung, skin, and liver) were carefully dissected under the microscope, with the exception of the blood which was collected by suction pipe from the heart directly. After washing in PBS 3 times to remove any residual blood contamination, the yolk sac, head, lung and skin were transferred to pre-warmed digestion medium containing 0.1 g/mL Collagenase I (Sigma, C2674, preheated to 37 °C) in RPMI 1640 medium. The samples were enzymatically digested at 37 °C in a humidified incubator for 20-30 minutes, with the samples being shaken periodically every five minutes until digested to single cell suspension. The liver was first cut into small pieces using scalpels, and then mechanically dissociated using syringes into single cell suspension, before removing the erythrocytes with lysis buffer (BD). The cells were then filtered through a 70 µm cell strainer after neutralizing enzymes. For the pediatric skin, the tissue was first incubated in RPMI 1640 with 1.2 U/ml dispase II (Roche, Indianapolis, IN) at 4 °C for 8 h prior to the separation of the epidermal layer. After separation, the epidermal layer was cut into pieces and digested with 0.25% trypsin-EDTA (Gibco) containing 0.25 mg/ml DNase I (Sigma DN25) for 30 min at 37°C in a humified incubator. After that, the epidermal layer was passed through a 70 µm cell strainer by grinding, and then washed with PBS. For bulk cultures, cells were sorted and co-cultured with MS5 for 14 days (100 cells per well). For the single cell cultures, single YSMPs were sorted and cultured in individual wells. Wells were observed under a light microscope and those with hematopoietic

	clusters (more than 50 round hematopoietic-like cells) at day 10 were counted as positive wells. At the end of the cultures, all cells including MS5 stromal cells in the selected positive wells were collected for further analyses with FACS to determine the lineage differentiation potential.
Instrument	BD ARIA 2
Software	BD FACSDIVA v8.0 and FlowJo X 10.0.7r2
Cell population abundance	The proportion of CD45+CD235a- cells in YS ranges from 10% to 30%. The cells, sorted by BD ARIA 2, were judged by morphology under microscope to assess their conditions, with the ones in good condition being picked by mouth pipette and directly placed into lysis buffer for single cell RNA library construction. The CD34+CD44+ population accounts for about 40% of CD45+ cells in YS.
Gating strategy	7AAD-CD45+CD235a-for STRT-seq; 7AAD- for 10x Genomics; 7AAD-CD45+CD34+CD44+ or 7AAD-CD45+CD34-CD44- for functional assay.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

Complement genes contribute sex-biased vulnerability in diverse disorders

https://doi.org/10.1038/s41586-020-2277-x

Received: 14 October 2019

Accepted: 28 February 2020

Published online: 11 May 2020

Check for updates

Nolan Kamitaki[1,2]✉, Aswin Sekar[1,2], Robert E. Handsaker[1,2], Heather de Rivera[1,2], Katherine Tooley[1,2], David L. Morris[3], Kimberly E. Taylor[4], Christopher W. Whelan[1,2], Philip Tombleson[3], Loes M. Olde Loohuis[5,6], Schizophrenia Working Group of the Psychiatric Genomics Consortium*, Michael Boehnke[7], Robert P. Kimberly[8], Kenneth M. Kaufman[9], John B. Harley[9], Carl D. Langefeld[10], Christine E. Seidman[1,11,12], Michele T. Pato[13], Carlos N. Pato[13], Roel A. Ophoff[5,6], Robert R. Graham[14], Lindsey A. Criswell[4], Timothy J. Vyse[3]✉ & Steven A. McCarroll[1,2]✉

Many common illnesses, for reasons that have not been identified, differentially affect men and women. For instance, the autoimmune diseases systemic lupus erythematosus (SLE) and Sjögren's syndrome affect nine times more women than men[1], whereas schizophrenia affects men with greater frequency and severity relative to women[2]. All three illnesses have their strongest common genetic associations in the major histocompatibility complex (MHC) locus, an association that in SLE and Sjögren's syndrome has long been thought to arise from alleles of the human leukocyte antigen (HLA) genes at that locus[3–6]. Here we show that variation of the complement component 4 (C4) genes *C4A* and *C4B*, which are also at the MHC locus and have been linked to increased risk for schizophrenia[7], generates 7-fold variation in risk for SLE and 16-fold variation in risk for Sjögren's syndrome among individuals with common C4 genotypes, with *C4A* protecting more strongly than *C4B* in both illnesses. The same alleles that increase risk for schizophrenia greatly reduce risk for SLE and Sjögren's syndrome. In all three illnesses, C4 alleles act more strongly in men than in women: common combinations of *C4A* and *C4B* generated 14-fold variation in risk for SLE, 31-fold variation in risk for Sjögren's syndrome, and 1.7-fold variation in schizophrenia risk among men (versus 6-fold, 15-fold and 1.26-fold variation in risk among women, respectively). At a protein level, both C4 and its effector C3 were present at higher levels in cerebrospinal fluid and plasma[8,9] in men than in women among adults aged between 20 and 50 years, corresponding to the ages of differential disease vulnerability. Sex differences in complement protein levels may help to explain the more potent effects of C4 alleles in men, women's greater risk of SLE and Sjögren's syndrome and men's greater vulnerability to schizophrenia. These results implicate the complement system as a source of sexual dimorphism in vulnerability to diverse illnesses.

SLE (commonly referred to as lupus) is a systemic autoimmune disease of unknown cause. Risk of SLE is largely (66%) heritable[10], although it may have environmental triggers, as onset often follows events that damage cells, such as infection and severe sunburn[11]. Most patients with SLE produce autoantibodies against nucleic acid complexes, including ribonucleoproteins and DNA[12].

In genetic studies, SLE is most strongly associated with variation across the MHC locus, which contains the HLA genes[3]. However, conclusive attribution of this association to specific genes and alleles has been difficult; the identities of the most likely genetic sources have been frequently revised as genetic studies have grown in size[4,5]. In several other autoimmune diseases, including type 1 diabetes, coeliac disease

[1]Department of Genetics, Harvard Medical School, Boston, MA, USA. [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Department of Medical and Molecular Genetics, King's College London, London, UK. [4]Rosalind Russell/Ephraim P. Engleman Rheumatology Research Center, Division of Rheumatology, UCSF School of Medicine, San Francisco, CA, USA. [5]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA. [6]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA. [7]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [8]Division of Clinical Immunology and Rheumatology, University of Alabama at Birmingham, Birmingham, AL, USA. [9]Center for Autoimmune Genomics and Etiology (CAGE), Department of Pediatrics, Cincinnati Children's Medical Center & University of Cincinnati and the US Department of Veterans Affairs Medical Center, Cincinnati, OH, USA. [10]Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston–Salem, NC, USA. [11]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [12]Cardiovascular Division, Brigham and Women's Hospital, Boston, MA, USA. [13]SUNY Downstate Medical Center, Brooklyn, NY, USA. [14]Human Genetics, Genentech, South San Francisco, CA, USA. *A list of participants and their affiliations appears in the online version of the paper. ✉e-mail: nolan_kamitaki@hms.harvard.edu; timothy.vyse@kcl.ac.uk; mccarroll@hms.harvard.edu

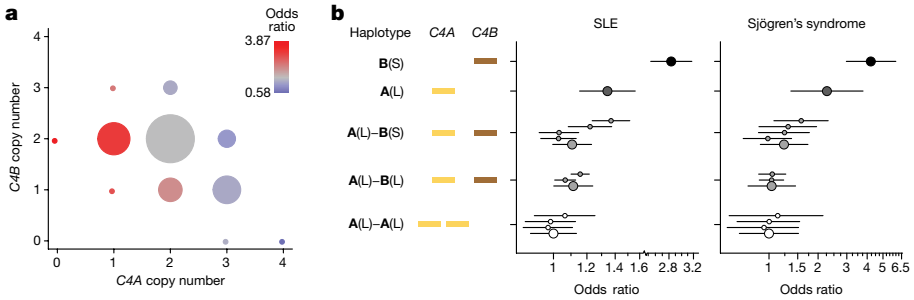


Fig. 1 | Association of SLE and Sjögren's syndrome with C4 alleles. a, Levels of SLE risk associated with 11 common combinations of *C4A* and *C4B* gene copy number. The colour of each circle reflects the level of SLE risk (odds ratio) associated with a specific combination of *C4A* and *C4B* gene copy numbers relative to the most common combination (two copies of *C4A* and two copies of *C4B*) in grey. The area of each circle is proportional to the number of individuals with that number of *C4A* and *C4B* genes. Paths from left to right on the plot reflect the effect of increasing *C4A* gene copy number (greatly reduced risk); paths from bottom to top reflect the effect of increasing *C4B* gene copy number (modestly reduced risk); and diagonal paths from upper left to lower right reflect the effect of exchanging *C4B* for *C4A* copies (modestly reduced risk). Data are from analysis of 6,748 patients with SLE and 11,516 unaffected controls of European ancestry. The odds ratios are reported with confidence intervals in Extended Data Fig. 2c. **b**, Risk of SLE and Sjögren's syndrome associated with common combinations of *C4A* and *C4B* gene copy number and flanking SNP haplotype. For each C4 locus structure, separate odds ratios are reported for each SNP haplotype background on which the C4 locus structure segregates. Data are from analyses of 6,748 patients with SLE and 11,516 controls (left) and 673 patients with Sjögren's syndrome and 1,153 controls (right). Error bars represent 95% confidence intervals around the effect-size estimate for each allele.

and rheumatoid arthritis, strong effects of the MHC locus arise from HLA alleles that cause the peptide-binding groove of HLA proteins to present a disease-critical autoantigen[13,14]. By contrast, in SLE, genetic variants in the MHC locus—including single nucleotide polymorphisms (SNPs) and HLA alleles—are broadly associated with the presence of diverse autoantibodies[15].

The *C4A* and *C4B* genes are also present in the MHC genomic region, between the class I and class II HLA genes. Classical complement proteins help eliminate debris from dead and damaged cells, attenuating the visibility of diverse intracellular proteins to the adaptive immune system. *C4A* and *C4B* commonly vary in genomic copy number[16] and encode complement proteins with distinct affinities for molecular targets[17,18]. SLE frequently presents with hypocomplementaemia that worsens during flares, possibly reflecting increased active consumption of complement[19]. Rare cases of severe, early-onset SLE can involve complete deficiency of a complement component (C4, C2 or C1Q)[20,21], and one of the strongest common-variant associations in SLE maps to *ITGAM*, which encodes a receptor for C3, the effector of C4 (ref. [22]). Although total C4 gene copy number is associated with SLE risk[23,24], this association is thought to arise from linkage disequilibrium (LD) with alleles of nearby HLA genes[25], which have been the focus of fine-mapping analyses[3,4].

The complex genetic variation of *C4A* and *C4B*—which consists of many alleles with different numbers of *C4A* and *C4B* genes—has been challenging to analyse in large cohorts. A recently feasible approach to this problem is based on imputation: people share long haplotypes with the same combinations of SNP and C4 alleles, such that *C4A* and *C4B* gene copy numbers can be imputed from SNP data[7]. To analyse *C4A* and *C4B* in large cohorts, we developed a way to identify C4 alleles from whole-genome sequence (WGS) data (Extended Data Fig. 1a, b), and then analysed WGS data from 1,265 individuals (from the Genomic Psychiatry Cohort[26,27]) to create a large multi-ancestry panel of 2,530 reference haplotypes of MHC-region SNPs, *C4A* alleles and *C4B* alleles (Extended Data Fig. 1c)—ten times as large as in earlier work[7]. We then analysed SNP data from the largest SLE genetic-association study[3] (ImmunoChip; 6,748 patients with SLE and 11,516 control subjects of European ancestry) (Extended Data Fig. 2a, b), imputing C4 alleles to estimate the SLE risk associated with common combinations of *C4A* and *C4B* gene copy numbers (Fig. 1a).

Groups of research participants with the eleven most common combinations of *C4A* and *C4B* gene copy number exhibited sevenfold variation in their relative risk of SLE (95% confidence interval (CI), [5.88, 8.61]; $P < 10^{-117}$ in total, Fig. 1a, Extended Data Fig. 2c). The relationship between SLE risk and C4 gene copy number exhibited consistent, logical patterns across the 11 genotype groups. For each *C4B* copy number, higher *C4A* copy number was associated with reduced SLE risk (Fig. 1a, Extended Data Fig. 2c). Conversely, for each *C4A* copy number, higher *C4B* copy number was associated with more modestly reduced SLE risk (Fig. 1a). Logistic-regression analysis estimated that the protection afforded by each copy of *C4A* (odds ratio 0.54; 95% confidence interval (CI): [0.51, 0.57]) was equivalent to that of 2.3 copies of *C4B* (odds ratio 0.77; 95% CI: [0.71, 0.82]). We calculated an initial C4 risk score as 2.3 times the number of *C4A* genes plus the number of *C4B* genes in an individual's genome. Despite clear limitations of this risk score—it is imperfectly imputed from flanking SNP haplotypes ($r^2 = 0.77$, Extended Data Table 1) and only approximates C4-derived risk by using a simple, linear model (to avoid overfitting the genetic data)—SNPs across the MHC genomic region tended to be associated with SLE in proportion to their level of LD with this risk score (Extended Data Fig. 3a).

Combinations of many different C4 alleles generate the observed variation in *C4A* and *C4B* gene copy number; particular *C4A* and *C4B* gene copy numbers have also arisen recurrently on multiple SNP haplotypes[7] (Extended Data Fig. 1c). Analysis of SLE risk in relation to each of these C4 alleles and SNP haplotypes reinforced the conclusion that *C4A* contributes strong protection, and *C4B* contributes more modest protection, from SLE, and that C4 genes (rather than nearby variants) are the principal drivers of this variation in risk levels (Fig. 1b).

These results prompted us to consider whether other autoimmune disorders with similar patterns of genetic association at the MHC genomic region might also be driven in part by variation of *C4A* and *C4B*. Primary Sjögren's syndrome is a heritable (54%)[28] systemic autoimmune disorder of exocrine glands, characterized primarily by dry eyes and mouth with other systemic effects. At a protein level, Sjögren's syndrome is (like SLE) characterized by diverse autoantibodies, including antinuclear antibodies targeting ribonucleoproteins[29], and hypocomplementaemia[30]. The largest source of common genetic risk for Sjögren's syndrome lies in the MHC genomic locus[31], with associations to the same haplotype(s) as in SLE[6] and with heterogeneous HLA associations in different ancestries[32]. We imputed C4 alleles into existing SNP data from a European-ancestry Sjögren's syndrome case–control cohort (673 cases and 1,153 controls). As in SLE, logistic-regression analyses found both *C4A* copy number (odds ratio 0.41; 95% CI: [0.34, 0.49]) and *C4B* copy number (OR: 0.67; 95% CI:

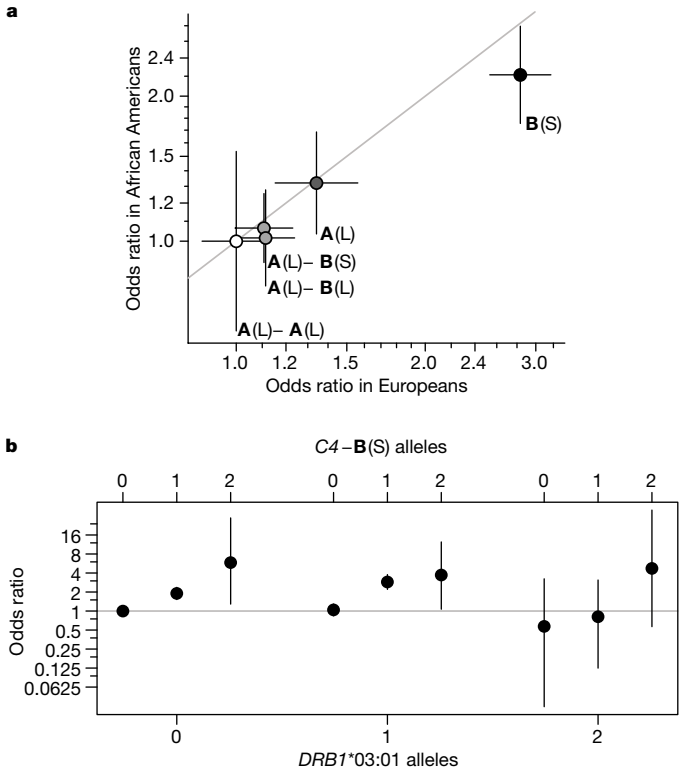


Fig. 2 | C4 and trans-ancestral analysis of the MHC-association signal in SLE. **a**, Common C4 alleles exhibit similar strengths of association to SLE (odds ratios) in European-ancestry and African American (1,494 SLE cases; 5,908 controls) cohorts. Error bars represent 95% confidence intervals around the effect size estimate for each sex. **b**, Analysis of SLE risk across combinations of C4-B(S) and *DRB1**03:01 genotypes in an African American SLE case–control cohort, in which the two alleles exhibit very little LD ($r^2 = 0.10$). On each *DRB1**03:01 genotype background, additional C4-B(S) alleles increase risk (that is, within each grouping). Whereas on each C4-B(S) background, *DRB1**03:01 alleles have no appreciable relationship with risk (this can be seen by comparing, for example, the first of the three points from each group). Error bars represent 95% confidence intervals around the effect-size estimate for each combination of C4-B(S) and *DRB1**03:01.

[0.53, 0.86]) to be protective against Sjögren's syndrome, generating a 16-fold variation in risk for Sjögren's syndrome (95% CI, [8.59, 30.89]; $P < 10^{-23}$ in total) among individuals with common C4 genotypes. The risk-equivalent ratio of *C4B* to *C4A* gene copies was similar in Sjögren's syndrome and SLE (about 2.3 to 1); furthermore, as with SLE, nearby SNPs associated with Sjögren's syndrome in proportion to their LD with a C4-derived risk score ((2.3)C4A + C4B) (Extended Data Fig. 3b), where C4A and C4B are the respective gene copy numbers. The distribution of Sjögren's syndrome risk across the individual *C4A* and *C4B* alleles and haplotypes revealed a pattern that, as in SLE, supported a greater protective effect from *C4A* than *C4B*, and little effect of flanking SNP haplotypes (Fig. 1b).

The association of SLE and Sjögren's syndrome with C4 gene copy number has long been attributed to the HLA-*DRB1*03:01* allele. In European populations, *DRB1*03:01* is in strong LD ($r^2 = 0.71$) with the common C4-B(S) allele, which lacks any *C4A* gene and is the highest-risk C4 allele in our analysis (Fig. 1b); many MHC-region SNPs associated with SLE and Sjögren's syndrome in proportion to their linkage-disequilibrium correlations with both C4 gene variation and *DRB1**03:01 (Extended Data Fig. 4a, b). Cohorts with other ancestries can have recombinant haplotypes that disambiguate the contributions of alleles that are in LD in Europeans. Among African Americans, we found that common C4 alleles exhibited far less LD with HLA alleles; in particular, the LD between C4-B(S) and *DRB1**03:01 was low ($r^2 = 0.10$) (Extended Data Table 2). Thus, genetic data from an African-American SLE cohort (1,494 cases and 5,908 controls) made it possible to distinguish between these potential genetic effects. Joint-association analysis of *C4A*, *C4B* and *DRB1*0301* implicated *C4A* ($P < 10^{-14}$) and *C4B* ($P < 10^{-5}$) but not *DRB1*0301* ($P = 0.29$) (Extended Data Table 3). Each C4 allele was associated with effect sizes of similar magnitude on SLE risk in Europeans and African Americans (Fig. 2a). An analysis specifically of combinations of C4-B(S) and *DRB1**03:01 allele dosages in African Americans showed that C4-B(S) alleles consistently increased SLE risk regardless of *DRB1**03:01 status, whereas *DRB1**03:01 had no consistent effect when controlling for C4-B(S) (Fig. 2b). Although C4 alleles had less LD with nearby variants on African American than on European haplotypes, SNPs across the genomic region associated with SLE in proportion to linkage-disequilibrium correlations with C4 variation in African Americans (Extended Data Fig. 4c).

Accounting for C4 alleles in jointly analysing the SLE-association data from African American and European ancestry cohorts also enabled mapping of an additional, more-modest genetic effect independent of *C4A* and *C4B*. This effect (tagged by rs2105898 and rs9271513) appeared to involve noncoding variation in the HLA class II XL9 region that is associated most strongly with expression levels (rather than the coding sequence) of many HLA class II genes (Extended Data Figs. 3c, d, 4d–l, 5 and Supplementary Note 1).

Alleles at C4 that increase dosage of *C4A* (and to a more modest extent *C4B*) appear to protect strongly against SLE and Sjögren's syndrome (Fig. 1a, b). By contrast, alleles that increase expression of *C4A* in the brain are more common among research participants with schizophrenia[6]. These same illnesses exhibit marked, and opposite, sex differences: SLE and Sjögren's syndrome are nine times more common among women of childbearing age than among men of a similar age[1], whereas in schizophrenia, women exhibit less severe symptoms, more frequent remission of symptoms, lower relapse rates and lower overall incidence[2]. Although the vast majority of genetic associations in complex diseases are shared between men and women[33], the SNPs most strongly associated with SLE risk within the MHC region are associated with larger potential effect sizes in men[34]. Thus, we sought to evaluate the possibility that the effects of C4 alleles on risk in SLE, Sjögren's syndrome and schizophrenia might differ between men and women.

Analysis indicated that the effects of C4 alleles were stronger in men. When a sex-by-C4 interaction term was included in association analyses, this term was significant for both SLE ($P = 0.002$) and schizophrenia ($P = 0.0024$), with larger C4 effects in men for both disorders. (Analysis of Sjögren's syndrome had limited power owing to the small number of men affected by Sjögren's syndrome). For both SLE and schizophrenia, the individual *C4A* and *C4B* alleles were consistently associated with stronger effects in men than women (Fig. 3a, b). SNPs across the MHC genomic region exhibited sex-biased association with SLE, Sjögren's syndrome and schizophrenia to the extent of their LD with C4 gene variation (Extended Data Fig. 6a–c).

The stronger effects of C4 alleles on male relative to female risk could arise from sex differences in C4 RNA expression, C4 protein levels or downstream responses to C4. Analysis of RNA expression in human tissues, using data from GTEx[35], identified no sex differences in C4 RNA expression in brain, blood, liver or lymphoblastoid cells (a more detailed description of this analysis can be found in Supplementary Note 2). We then analysed C4 protein in cerebrospinal fluid (CSF) from two panels of adult research participants ($n = 589$ total) in whom we had also measured C4 gene copy number (by direct genotyping or imputation). CSF C4 protein levels correlated strongly with C4 gene copy number ($P < 10^{-10}$, Extended Data Fig. 7a), so we normalized C4 protein measurements to the number of C4 gene copies. CSF from adult men contained on average 27% more C4 protein per C4 gene copy than CSF from women (meta-analysis $P = 9.9 \times 10^{-6}$, Fig. 3c). C4 acts by activating the complement component 3 (C3) protein, promoting C3 deposition

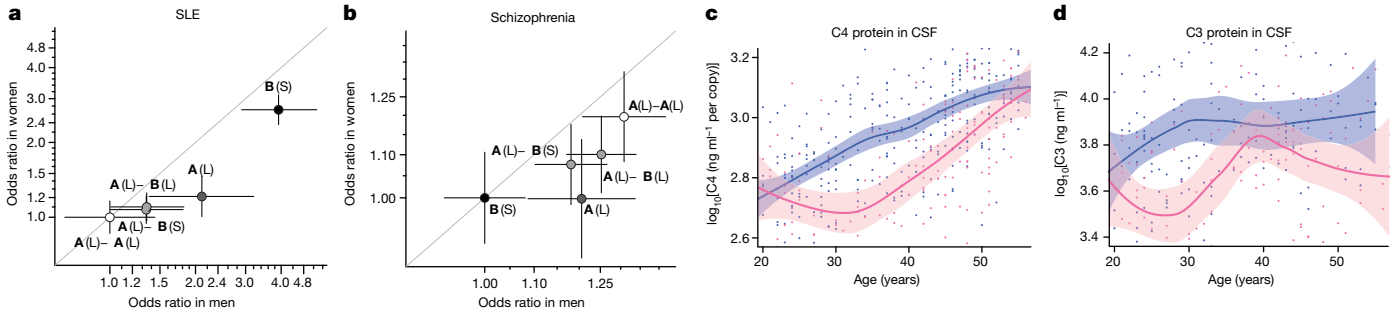


Fig. 3 | Sex differences in the magnitude of C4 genetic effects and complement protein concentrations. a, SLE risk (odds ratios) associated with the four most common C4 alleles in men (*x* axis) and women (*y* axis) among 6,748 affected and 11,516 unaffected individuals of European ancestry. For each sex, the lowest-risk allele (C4-A(L)-A(L)) is used as a reference (odds ratio of 1.0). Shading of each point reflects the relative level of SLE risk (darker indicates greater risk) conferred by *C4A* and *C4B* copy numbers as in Fig. 2b. Error bars represent 95% confidence intervals around the effect size estimate for each sex. **b**, Schizophrenia risk (odds ratios) associated with the four most common C4 alleles in men (*x* axis) and women (*y* axis) among 28,799 affected and 35,986 unaffected individuals of European ancestry, aggregated by the Psychiatric Genomics Consortium[43]. For each sex, the lowest-risk allele (C4-B(S)) is used as a reference (odds ratio of 1.0). For visual comparison with **a**, shading of each allele reflects the relative level of SLE risk. Error bars represent 95% confidence intervals around the effect size estimate for each sex. **c**, Concentrations of C4 protein in CSF sampled from 340 adult men (blue) and 167 adult women (pink) as a function of age with locally estimated scatterplot smoothing (LOESS). Concentrations are normalized to the number of C4 gene copies in an individual's genome (a strong independent source of variance, Extended Data Fig. 7a) and shown on a $\log_{10}$ scale as a LOESS curve. Shaded regions represent 95% confidence intervals derived during LOESS. **d**, Levels of C3 protein in CSF from 179 adult men and 125 adult women as a function of age. Concentrations are shown on a $\log_{10}$ scale as a LOESS curve. Shaded regions represent 95% confidence intervals derived during LOESS.

onto targets in tissues. CSF levels of C3 protein were also on average 42% higher among men than women (meta-analysis $P = 7.5 \times 10^{-7}$, Fig. 3d).

The elevated concentrations of C3 and C4 proteins in CSF of men parallel earlier findings showing that, in plasma, C3 and C4 are also present at higher levels in men than women[8,9]. The large sample size ($n > 50{,}000$) of the plasma studies enables sex differences to be further analysed as a function of age. Both men and women undergo age-dependent elevation of C4 and C3 levels in plasma, but this occurs early in adulthood (20–30 years of age) in men and closer to menopause (40–50 years of age) in women, with the result that male–female differences in complement protein levels are observed primarily during the reproductive years (20–50 years of age)[8,9]. We replicated these findings using measurements of C3 and gene copy number-corrected C4 protein in plasma from adults, finding (as in the earlier plasma studies[8,9] and in CSF; Fig. 3c, d) that these differences are most pronounced during the reproductively active years of adulthood (20–50 years of age) (Extended Data Fig. 7b–d). We also observed that patients with Sjögren's syndrome have lower C4 serum levels than unaffected individuals ($P < 1\text{x}10^{-20}$, Extended Data Fig. 7e) even after correcting for C4 gene copy number ($P < 1\text{x}10^{-8}$, Extended Data Fig. 7f), suggesting that hypocomplementaemia in Sjögren's syndrome is not simply due to C4 genetics but also reflects disease effects on background complement levels, for example, owing to complement consumption. The ages of pronounced sex difference in complement levels correspond with the ages at which men and women differ in disease incidence: in schizophrenia, men outnumber women among cases incident in early adulthood, but not among cases incident after 40 years of age[2]; in SLE, women greatly outnumber men among cases incident during the child-bearing years, but not among cases incident after 50 years of age or during childhood[36]; in Sjögren's syndrome, the high relative vulnerability of women declines in magnitude after 50 years of age[37].

Our results indicate that the MHC genomic region shapes vulnerability in lupus and Sjögren's syndrome—two of the three most common rheumatic autoimmune diseases—in a very different way than in type I diabetes, rheumatoid arthritis and coeliac disease. In those diseases, precise interactions between HLA protein variants and specific autoantigens determine risk[13,14]. In SLE and Sjögren's syndrome, however, the genetic variation implicated here points instead to the continuous, chronic interaction of the immune system with a large number of potential autoantigens. Because complement facilitates the rapid clearance of debris from dead and injured cells, increased levels of C4 protein probably attenuate interactions between the adaptive immune system and ribonuclear self-antigens at sites of cell injury, pre-empting the development of autoimmunity. The additional C4-independent genetic risk effect described here (associated with rs2105898) may also affect autoimmunity broadly, rather than in an antigen-specific manner, by regulating expression of many HLA class II genes (including *DRB1*, *DQA1* and *DQB1*). Mouse models of SLE indicate that once tolerance is broken for one self-antigen, autoreactive germinal centres generate B cells targeting other self-antigens[38]; such 'epitope spreading' could lead to autoreactivity against many related autoantigens, regardless of which antigen(s) are involved in the earliest interactions with immune cells. Further supporting such a model, higher copy number of C4 is associated with lower risk of AQP4-IgG-seropositive neuromyelitis optica[39], in which seropositive patients have increased incidence of other non-organ-specific autoantibodies such as those seen in SLE and Sjögren's syndrome[40]. B cells also express the complement receptors CR1 and CR2[41], providing an additional candidate mechanism for regulation by C4 and C3.

We note that the role of complement proteins in preventing the emergence of autoimmunity may be very different than their (potentially disease-exacerbating) role once autoimmunity has been established. Also, our genetic findings address the development of SLE and Sjögren's syndrome rather than complications that arise in any specific organ. A few per cent of patients with SLE develop neurological complications that can include psychosis[42]; although psychosis is also a symptom of schizophrenia, neurological complications of SLE do not resemble schizophrenia more broadly, and probably have a different aetiology.

The same C4 alleles that increase vulnerability to schizophrenia appeared to protect strongly against SLE and Sjögren's syndrome. This pleiotropy will need to be considered in efforts to engage the complement system therapeutically. The complement system contributed to these pleiotropic effects more strongly in men than in women. Moreover, though the natural allelic series at C4 enabled human-genetic analysis to establish dose–risk relationships for C4 in men and women, sexual dimorphism in the levels of complement protein also included complement component 3 (C3). Why and how this sexual dimorphism in the complement system has evolved in

humans poses interesting questions for immune and evolutionary biology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2277-x.

1. Ngo, S. T., Steyn, F. J. & McCombe, P. A. Gender differences in autoimmune disease. *Front. Neuroendocrinol.* **35**, 347–369 (2014).
2. Abel, K. M., Drake, R. & Goldstein, J. M. Sex differences in schizophrenia. *Int. Rev. Psychiatry* **22**, 417–428 (2010).
3. Langefeld, C. D. et al. Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* **8**, 16021 (2017).
4. Rioux, J. D. et al. Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proc. Natl Acad. Sci. USA* **106**, 18680–18685 (2009).
5. Hanscombe, K. B. et al. Genetic fine mapping of systemic lupus erythematosus MHC associations in Europeans and African Americans. *Hum. Mol. Genet.* **27**, 3813–3824 (2018).
6. Cruz-Tapias, P., Rojas-Villarraga, A., Maier-Moore, S. & Anaya, J. M. HLA and Sjögren's syndrome susceptibility. a meta-analysis of worldwide studies. *Autoimmun. Rev.* **11**, 281–287 (2012).
7. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
8. Gaya da Costa, M. et al. Age and sex-associated changes of complement activity and complement levels in a healthy Caucasian population. *Front. Immunol.* **9**, 2664 (2018).
9. Ritchie, R. F. et al. Reference distributions for complement proteins C3 and C4: a practical, simple and clinically relevant approach in a large cohort. *J. Clin. Lab. Anal.* **18**, 1–8 (2004).
10. Lawrence, J. S., Martins, C. L. & Drake, G. L. A family survey of lupus erythematosus. 1. Heritability. *J. Rheumatol.* **14**, 913–921 (1987).
11. Lipsky, P. E. Systemic lupus erythematosus: an autoimmune disease of B cell hyperactivity. *Nat. Immunol.* **2**, 764–766 (2001).
12. Ippolito, A. et al. Autoantibodies in systemic lupus erythematosus: comparison of historical and current assessment of seropositivity. *Lupus* **20**, 250–255 (2011).
13. Lee, K. H., Wucherpfennig, K. W. & Wiley, D. C. Structure of a human insulin peptide–HLA–DQ8 complex and susceptibility to type 1 diabetes. *Nat. Immunol.* **2**, 501–507 (2001).
14. Raychaudhuri, S. et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291–296 (2012).
15. Morris, D. L. et al. MHC associations with clinical and autoantibody manifestations in European SLE. *Genes Immun.* **15**, 210–217 (2014).
16. Bánlaki, Z., Doleschall, M., Rajczy, K., Fust, G. & Szilágyi, A. Fine-tuned characterization of RCCX copy number variants and their relationship with extended MHC haplotypes. *Genes Immun.* **13**, 530–535 (2012).
17. Isenman, D. E. & Young, J. R. The molecular basis for the difference in immune hemolysis activity of the Chido and Rodgers isotypes of human complement component C4. *J. Immunol.* **132**, 3019–3027 (1984).
18. Law, S. K., Dodds, A. W. & Porter, R. R. A comparison of the properties of two classes, C4A and C4B, of the human complement component C4. *EMBO J.* **3**, 1819–1823 (1984).
19. Birmingham, D. J. et al. The complex nature of serum C3 and C4 as biomarkers of lupus renal flare. *Lupus* **19**, 1272–1280 (2010).
20. Ross, S. C. & Densen, P. Complement deficiency states and infection: epidemiology, pathogenesis and consequences of neisserial and other infections in an immune deficiency. *Medicine* **63**, 243–273 (1984).
21. Wu, Y. L., Hauptmann, G., Viguier, M. & Yu, C. Y. Molecular basis of complete complement C4 deficiency in two North-African families with systemic lupus erythematosus. *Genes Immun.* **10**, 433–445 (2009).
22. International Consortium for Systemic Lupus Erythematosus. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM, PXK, KIAA1542* and other loci. *Nat. Genet.* **40**, 204–210 (2008).
23. Yang, Y. et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am. J. Hum. Genet.* **80**, 1037–1054 (2007).
24. Jüptner, M. et al. Low copy numbers of complement C4 and homozygous deficiency of C4A may predispose to severe disease and earlier disease onset in patients with systemic lupus erythematosus. *Lupus* **27**, 600–609 (2018).
25. Boteva, L. et al. Genetically determined partial complement C4 deficiency states are not independent risk factors for SLE in UK and Spanish populations. *Am. J. Hum. Genet.* **90**, 445–456 (2012).
26. Pato, M. T. et al. The genomic psychiatry cohort: partners in discovery. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **162**, 306–312 (2013).
27. Sanders, S. J. et al. Whole genome sequencing in psychiatric disorders: the WGSPD consortium. *Nat. Neurosci.* **20**, 1661–1668 (2017).
28. Kuo, C. F. et al. Familial risk of Sjögren's syndrome and co-aggregation of autoimmune diseases in affected families: a nationwide population study. *Arthritis Rheumatol.* **67**, 1904–1912 (2015).
29. Fayyaz, A., Kurien, B. T. & Scofield, R. H. Autoantibodies in Sjögren's Syndrome. *Rheum. Dis. Clin. North Am.* **42**, 419–434 (2016).
30. Ramos-Casals, M. et al. Hypocomplementaemia as an immunological marker of morbidity and mortality in patients with primary Sjögren's syndrome. *Rheumatology* **44**, 89–94 (2005).
31. Chused, T. M., Kassan, S. S., Opelz, G., Moutsopoulos, H. M. & Terasaki, P. I. Sjögren's syndrome association with HLA-Dw3. *N. Engl. J. Med.* **296**, 895–897 (1977).
32. Taylor, K. E. et al. Genome-wide association analysis reveals genetic heterogeneity of Sjögren's syndrome according to ancestry. *Arthritis Rheumatol.* **69**, 1294–1305 (2017).
33. Khramtsova, E. A., Davis, L. K. & Stranger, B. E. The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.* **20**, 173–190 (2019).
34. Hughes, T. et al. Analysis of autosomal genes reveals gene–sex interactions and higher total genetic risk in men with systemic lupus erythematosus. *Ann. Rheum. Dis.* **71**, 694–699 (2012).
35. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
36. Brinks, R. et al. Age-specific and sex-specific incidence of systemic lupus erythematosus: an estimate from cross-sectional claims data of 2.3 million people in the German statutory health insurance 2002. *Lupus Sci. Med.* **3**, e000181 (2016).
37. Kim, H. J. et al. Incidence, mortality, and causes of death in physician-diagnosed primary Sjögren's syndrome in Korea: A nationwide, population-based study. *Semin. Arthritis Rheum.* **47**, 222–227 (2017).
38. Degn, S. E. et al. Clonal evolution of autoreactive germinal centers. *Cell* **170**, 913–926 (2017).
39. Estrada, K. et al. A whole-genome sequence study identifies genetic risk factors for neuromyelitis optica. *Nat. Commun.* **9**, 1929 (2018).
40. Pittock, S. J. et al. Neuromyelitis optica and non organ-specific autoimmunity. *Arch. Neurol.* **65**, 78–83 (2008).
41. Erdei, A. et al. Expression and role of CR1 and CR2 on B and T lymphocytes under physiological and autoimmune conditions. *Mol. Immunol.* **46**, 2767–2773 (2009).
42. Unterman, A. et al. Neuropsychiatric syndromes in systemic lupus erythematosus: a meta-analysis. *Semin. Arthritis Rheum.* **41**, 1–11 (2011).
43. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

Schizophrenia Working Group of the Psychiatric Genomics Consortium

Stephan Ripke[15,16], Benjamin M. Neale[15,16,17,18], Aiden Corvin[19], James T. R. Walters[20], Kai-How Farh[15], Peter A. Holmans[20,21], Phil Lee[15,16,18], Brendan Bulik-Sullivan[15,16], David A. Collier[22,23], Hailiang Huang[15,17], Tune H. Pers[17,24,25], Ingrid Agartz[26,27,28], Esben Agerbo[29,30,31], Margot Albus[32], Madeline Alexander[33], Farooq Amin[34,35], Silviu A. Bacanu[36], Martin Begemann[37], Richard A. Belliveau Jr[16], Judit Bene[38,39], Sarah E. Bergen[16,40], Elizabeth Bevilacqua[16], Tim B. Bigdeli[36], Donald W. Black[41], Richard Bruggeman[42], Nancy G. Buccola[43], Randy L. Buckner[44,45,46], William Byerley[47], Wiepke Cahn[48], Guiqing Cai[49,50], Murray J. Cairns[51,52,53], Dominique Campion[54], Rita M. Cantor[55], Vaughan J. Carr[51,56], Noa Carrera[20], Stanley V. Catts[51,57], Kimberly D. Chambert[16], Raymond C. K. Chan[58], Ronald Y. L. Chen[59], Eric Y. H. Chen[59,60], Wei Cheng[61], Eric F. C. Cheung[62], Siow Ann Chong[63], C. Robert Cloninger[64], David Cohen[65], Nadine Cohen[66], Paul Cormican[19], Nick Craddock[20,21], Benedicto Crespo-Facorro[67], James J. Crowley[68], David Curtis[69,70], Michael Davidson[71], Kenneth L. Davis[50], Franziska Degenhardt[72,73], Jurgen Del Favero[74], Lynn E. DeLisi[75,76], Ditte Demontis[31,77,78], Dimitris Dikeos[79], Timothy Dinan[80], Srdjan Djurovic[28,81], Gary Donohoe[19,82], Elodie Drapeau[50], Jubao Duan[83,84], Frank Dudbridge[85], Naser Durmishi[86], Peter Eichhammer[87], Johan Eriksson[88,89,90], Valentina Escott-Price[20], Laurent Essioux[91], Ayman H. Fanous[92,93,94,95], Martilias S. Farrell[68], Josef Frank[96], Lude Franke[97], Robert Freedman[98], Nelson B. Freimer[99], Marion Friedl[100], Joseph I. Friedman[50], Menachem Fromer[15,16,18,101], Giulio Genovese[16], Lyudmila Georgieva[20], Elliot S. Gershon[102], Ina Giegling[100,103], Paola Giusti-Rodríguez[68], Stephanie Godard[104], Jacqueline I. Goldstein[15,17], Vera Golimbet[105], Srihari Gopal[106], Jacob Gratten[107], Lieuwe de Haan[108], Marina Mitjans[37], Marian L. Hamshere[20], Mark Hansen[109], Thomas Hansen[31,110], Vahram Haroutunian[50,111,112], Annette M. Hartmann[100], Frans A. Henskens[51,113,114], Stefan Herms[72,73,115], Joel N. Hirschhorn[17,25,116], Per Hoffmann[72,73,115], Andrea Hofman[72,73], Mads V. Hollegaard[117], David M. Hougaard[117], Masashi Ikeda[118], Inge Joa[119], Antonio Julià[120], René S. Kahn[48], Luba Kalaydjieva[121], Sena Karachanak-Yankova[122], Juha Karjalainen[97], David Kavanagh[20], Matthew C. Keller[123], Brian J. Kelly[52], James L. Kennedy[124,125,126], Andrey Khrunin[127], Yunjung Kim[68], Janis Klovins[128], James A. Knowles[129], Bettina Konte[100], Vaidutis Kucinskas[130], Zita Ausrele Kucinskiene[130], Hana Kuzelova-Ptackova[131], Anna K. Kähler[40], Claudine Laurent[33,132], Jimmy Lee Chee Keong[63,133], S. Hong Lee[107], Sophie E. Legge[20], Bernard Lerer[134], Miaoxin Li[59,60,135], Tao Li[136], Kung-Yee Liang[137], Jeffrey Lieberman[138], Svetlana Limborska[127], Carmel M. Loughland[51,52], Jan Lubinski[139], Jouko Lönnqvist[140], Milan Macek Jr[131], Patrik K. E. Magnusson[40], Brion S. Maher[141], Wolfgang Maier[142], Jacques Mallet[143], Sara Marsal[120], Manuel Mattheisen[31,77,78,144], Morten Mattingsdal[28,145], Robert W. McCarley[75,76], Colm McDonald[146], Andrew M. McIntosh[147,148], Sandra Meier[96], Carin J. Meijer[108], Bela Melegh[38,39], Ingrid Melle[28,149], Raquelle I. Mesholam-Gately[75,150], Andres Metspalu[151], Patricia T. Michie[51,152], Lili Milani[151], Vihra Milanova[153], Younes Mokrab[22], Derek W. Morris[19,82], Ole Mors[31,77,154], Kieran C. Murphy[155], Robin M. Murray[156], Inez Myin-Germeys[157], Bertram Müller-Myhsok[158,159,160], Mari Nelis[151], Igor Nenadic[161], Deborah A. Nertney[162], Gerald Nestadt[163], Kristin K. Nicodemus[164], Liene Nikitina-Zake[128], Laura Nisenbaum[165], Annelie Nordin[166], Eadbhard O'Callaghan[167], Colm O'Dushlaine[16], F. Anthony O'Neill[168], Sang-Yun Oh[169], Ann Olincy[98], Line Olsen[31,110], Jim Van Os[157,170], Psychosis Endophenotypes International Consortium*, Christos Pantelis[51,171], George N. Papadimitriou[79], Agnes A. Steixner[37], Elena Parkhomenko[50], Michele T. Pato[129], Tiina Paunio[172,173], Milica Pejovic-Milovancevic[174], Diana O. Perkins[175], Olli Pietiläinen[173,176], Jonathan Pimm[70], Andrew J. Pocklington[20], John Powell[156], Alkes Price[17,177], Ann E. Pulver[163], Shaun M. Purcell[101], Digby Quested[178], Henrik B. Rasmussen[31,110], Abraham Reichenberg[50], Mark A. Reimers[179], Alexander L. Richards[20], Joshua L. Roffman[44,46], Panos Roussos[101,180], Douglas M. Ruderfer[20,101], Veikko Salomaa[90], Alan R. Sanders[83,84], Ulrich Schall[51,52], Christian R. Schubert[181], Thomas G. Schulze[96,182], Sibylle G. Schwab[183], Edward M. Scolnick[16], Rodney J. Scott[51,53,184], Larry J. Seidman[75,150], Jianxin Shi[185], Engilbert Sigurdsson[186], Teimuraz Silagadze[187], Jeremy M. Silverman[50,188], Kang Sim[63], Petr Slominsky[127], Jordan W. Smoller[16,18], Hon-Cheong So[59], Chris C. A. Spencer[189], Eli A. Stahl[17,101], Hreinn Stefansson[190], Stacy Steinberg[190], Elisabeth Stogmann[191], Richard E. Straub[192], Eric Strengman[48,193], Jana Strohmaier[96], T. Scott Stroup[138], Mythily Subramaniam[63], Jaana Suvisaari[140], Dragan M. Svrakic[64], Jin P. Szatkiewicz[68], Erik Söderman[26], Srinivas Thirumalai[194], Draga Toncheva[122], Paul A. Tooney[51,52,53], Sarah Tosato[195], Juha Veijola[196,197], John Waddington[198], Dermot Walsh[199], Dai Wang[106], Qiang Wang[136], Bradley T. Webb[36], Mark Weiser[71], Dieter B. Wildenauer[200], Nigel M. Williams[20], Stephanie Williams[68], Stephanie H. Witt[96], Aaron R. Wolen[179], Emily H. M. Wong[59], Brandon K. Wormley[36], Jing Qin Wu[51,53], Hualin Simon Xi[201], Clement C. Zai[124,125], Xuebin Zheng[202], Fritz Zimprich[191], Naomi R. Wray[107], Kari Stefansson[190], Peter M. Visscher[107], Wellcome Trust Case–Control Consortium 2*, Rolf Adolfsson[166], Ole A. Andreassen[28,149], Douglas H. R. Blackwood[148], Elvira Bramon[203], Joseph D. Buxbaum[49,50,111,204], Anders D. Børglum[31,77,78,154], Sven Cichon[72,73,115,205], Ariel Darvasi[206], Enrico Domenici[207], Hannelore Ehrenreich[37], Tõnu Esko[17,25,116,151], Pablo V. Gejman[83,84], Michael Gill[19], Hugh Gurling[70], Christina M. Hultman[40], Nakao Iwata[118], Assen V. Jablensky[51,200,208,209], Erik G. Jönsson[26,28], Kenneth S. Kendler[210], George Kirov[20], Jo Knight[124,125,126], Todd Lencz[211,212,213], Douglas F. Levinson[33], Qingqin S. Li[106], Jianjun Liu[202,214], Anil K. Malhotra[211,212,213], Steven A. McCarroll[16,116], Andrew McQuillin[70], Jennifer L. Moran[16], Preben B. Mortensen[29,30,31], Bryan J. Mowry[107,215], Markus M. Nöthen[72,73], Roel A. Ophoff[48,55,99], Michael J. Owen[20,21], Aarno Palotie[16,18,176], Carlos N. Pato[129], Tracey L. Petryshen[16,75,216], Danielle Posthuma[217,218,219], Marcella Rietschel[96], Brien P. Riley[210], Dan Rujescu[100,103], Pak C. Sham[59,60,135], Pamela Sklar[101,111,180], David St Clair[220], Daniel R. Weinberger[192,221], Jens R. Wendland[181], Thomas Werge[31,110,222], Mark J. Daly[15,16,17], Patrick F. Sullivan[40,68,175] & Michael C. O'Donovan[20,21]

Psychosis Endophenotype International Consortium

Maria J. Arranz[170,223], Steven Bakker[48], Stephan Bender[224,225], Elvira Bramon[170,226,227], David A. Collier[22,23], Benedicto Crespo-Facorro[228,229], Jeremy Hall[148], Conrad Iyegbe[170], Assen V. Jablensky[230], René S. Kahn[48], Luba Kalaydjieva[121,231], Stephen Lawrie[148], Cathryn M. Lewis[170], Kuang Lin[170], Don H. Linszen[232], Ignacio Mata[228,229], Andrew M. McIntosh[148], Robin M. Murray[156], Roel A. Ophoff[99], Jim Van Os[157,170], John Powell[170], Dan Rujescu[100,103], Muriel Walshe[170], Matthias Weisbrod[225] & Durk Wiersma[233]

Wellcome Trust Case–Control Consortium 2

Peter Donnelly[192,234], Ines Barroso[235], Jenefer M. Blackwell[236,237], Elvira Bramon[205], Matthew A. Brown[238], Juan P. Casas[239,240], Aiden Corvin[19], Panos Deloukas[235], Audrey Duncanson[241], Janusz Jankowski[242], Hugh S. Markus[243], Christopher G. Mathew[244], Colin N. A. Palmer[245], Robert Plomin[23], Anna Rautanen[192], Stephen J. Sawcer[246], Richard C. Trembath[244], Ananth C. Viswanathan[247,248], Nicholas W. Wood[249], Chris C. A. Spencer[192], Gavin Band[192], Céline Bellenguez[192], Peter Donnelly[192,234], Colin Freeman[192], Eleni Giannoulatou[192], Garrett Hellenthal[192], Richard Pearson[192], Matti Pirinen[192], Amy Strange[192], Zhan Su[192], Damjan Vukcevic[192], Cordelia Langford[235], Ines Barroso[235], Hannah Blackburn[235], Suzannah J. Bumpstead[235], Panos Deloukas[235], Serge Dronov[235], Sarah Edkins[235], Matthew Gillman[235], Emma Gray[235], Rhian Gwilliam[235], Naomi Hammond[235], Sarah E. Hunt[235], Alagurevathi Jayakumar[235], Jennifer Liddle[235], Owen T. McCann[235], Simon C. Potter[235], Radhi Ravindrarajah[235], Michelle Ricketts[235], Avazeh Tashakkori-Ghanbaria[235], Matthew Waller[235], Paul Weston[235], Pamela Whittaker[235], Sara Widaa[235], Christopher G. Mathew[244], Jenefer M. Blackwell[236,237], Matthew A. Brown[238], Aiden Corvin[19], Mark I. McCarthy[250] & Chris C. A. Spencer[192]

[15]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [16]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [17]Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [18]Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [19]Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity College Dublin, Dublin, Ireland. [20]MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK. [21]National Centre for Mental Health, Cardiff University, Cardiff, UK. [22]Eli Lilly, Windlesham, UK. [23]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, UK. [24]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark. [25]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA, USA. [26]Department of Clinical Neuroscience, Psychiatry Section, Karolinska Institutet, Stockholm, Sweden. [27]Department of Psychiatry, Diakonhjemmet Hospital, Oslo, Norway. [28]NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [29]Centre for Integrative Register-based Research, CIRRAU, Aarhus University, Aarhus, Denmark. [30]National Centre for Register-based Research, Aarhus University, Aarhus, Denmark. [31]The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Aarhus, Denmark. [32]State Mental Hospital, Haar, Germany. [33]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA. [34]Department of Psychiatry and Behavioral Sciences, Atlanta Veterans Affairs Medical Center, Atlanta, GA, USA. [35]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, GA, USA. [36]Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA. [37]Clinical Neuroscience, Max Planck Institute of Experimental Medicine, Göttingen, Germany. [38]Department of Medical Genetics, University of Pécs, Pécs, Hungary. [39]Szentagothai Research Center, University of Pécs, Pécs, Hungary. [40]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [41]Department of Psychiatry, University of Iowa Carver College of Medicine, Iowa City, IA, USA. [42]Department of Psychiatry, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [43]School of Nursing, Louisiana State University Health Sciences Center, New Orleans, LA, USA. [44]Athinoula A. Martinos Center, Massachusetts General Hospital, Boston, MA, USA. [45]Center for Brain Science, Harvard University, Cambridge, MA, USA. [46]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [47]Department of Psychiatry, University of California at San Francisco, San Francisco, CA, USA. [48]Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, The Netherlands. [49]Department of Human Genetics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [50]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [51]Schizophrenia Research Institute, Sydney, New South Wales, Australia. [52]Priority Centre for Translational Neuroscience and Mental Health, University of Newcastle, Newcastle, New South Wales, Australia. [53]School of Biomedical Sciences and Pharmacy, University of Newcastle, Callaghan, New South Wales, Australia. [54]Centre Hospitalier du Rouvray and INSERM U1079 Faculty of Medicine, Rouen, France. [55]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA. [56]School of Psychiatry, University of New South Wales, Sydney, New South Wales, Australia. [57]Royal Brisbane and Women's Hospital, University of Queensland, Brisbane, Australia. [58]Institute of Psychology, Chinese Academy of Science, Beijing, China. [59]Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [60]State Key Laboratory for Brain and Cognitive Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [61]Department of

Computer Science, University of North Carolina, Chapel Hill, NC, USA. [62]Castle Peak Hospital, Hong Kong, China. [63]Institute of Mental Health, Singapore, Singapore. [64]Department of Psychiatry, Washington University, St Louis, MO, USA. [65]Department of Child and Adolescent Psychiatry, Assistance Publique Hopitaux de Paris, Pierre and Marie Curie Faculty of Medicine and Institute for Intelligent Systems and Robotics, Paris, France. [66]Blue Note Biosciences, Princeton, NJ, USA. [67]University Hospital Marqués de Valdecilla, Instituto de Formación e Investigación Marqués de Valdecilla, University of Cantabria, Santander, Spain. [68]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [69]Department of Psychological Medicine, Queen Mary University of London, London, UK. [70]Molecular Psychiatry Laboratory, Division of Psychiatry, University College London, London, UK. [71]Sheba Medical Center, Tel Hashomer, Israel. [72]Department of Genomics, Life and Brain Center, Bonn, Germany. [73]Institute of Human Genetics, University of Bonn, Bonn, Germany. [74]Applied Molecular Genomics Unit, VIB Department of Molecular Genetics, University of Antwerp, Antwerp, Belgium. [75]Department of Psychiatry, Harvard Medical School, Boston, MA, USA. [76]VA Boston Health Care System, Brockton, MA, USA. [77]Centre for Integrative Sequencing, iSEQ, Aarhus University, Aarhus, Denmark. [78]Department of Biomedicine, Aarhus University, Aarhus, Denmark. [79]First Department of Psychiatry, University of Athens Medical School, Athens, Greece. [80]Department of Psychiatry, University College Cork, Cork, Ireland. [81]Department of Medical Genetics, Oslo University Hospital, Oslo, Norway. [82]Cognitive Genetics and Therapy Group, School of Psychology and Discipline of Biochemistry, National University of Ireland Galway, Galway, Ireland. [83]Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL, USA. [84]Department of Psychiatry and Behavioral Sciences, NorthShore University HealthSystem, Evanston, IL, USA. [85]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [86]Department of Child and Adolescent Psychiatry, University Clinic of Psychiatry, Skopje, Republic of Macedonia. [87]Department of Psychiatry, University of Regensburg, Regensburg, Germany. [88]Department of General Practice, Helsinki University Central Hospital, University of Helsinki, Helsinki, Finland. [89]Folkhälsan Research Center, Helsinki, Finland, Biomedicum Helsinki 1, Helsinki, Finland. [90]National Institute for Health and Welfare, Helsinki, Finland. [91]Translational Technologies and Bioinformatics, Pharma Research and Early Development, F. Hoffman-La Roche, Switzerland. [92]Department of Psychiatry, Georgetown University School of Medicine, Washington, DC, USA. [93]Department of Psychiatry, Keck School of Medicine of the University of Southern California, Los Angeles, CA, USA. [94]Department of Psychiatry, Virginia Commonwealth University School of Medicine, Richmond, VA, USA. [95]Mental Health Service Line, Washington VA Medical Center, Washington, DC, USA. [96]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Heidelberg, Germany. [97]Department of Genetics, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands. [98]Department of Psychiatry, University of Colorado Denver, Aurora, CO, USA. [99]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA. [100]Department of Psychiatry, University of Halle, Halle, Germany. [101]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [102]Departments of Psychiatry and Human Genetics, University of Chicago, Chicago, IL, USA. [103]Department of Psychiatry, University of Munich, Munich, Germany. [104]Departments of Psychiatry and Human and Molecular Genetics, INSERM, Institut de Myologie, Hôpital de la Pitiè-Salpêtrière, Paris, France. [105]Mental Health Research Centre, Russian Academy of Medical Sciences, Moscow, Russia. [106]Neuroscience Therapeutic Area, Janssen Research and Development, Raritan, NJ, USA. [107]Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia. [108]Department of Psychiatry, Academic Medical Centre University of Amsterdam, Amsterdam, The Netherlands. [109]Illumina, La Jolla, CA, USA. [110]Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Mental Health Services Copenhagen, Copenhagen, Denmark. [111]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [112]J. J. Peters VA Medical Center, Bronx, New York, NY, USA. [113]Priority Research Centre for Health Behaviour, University of Newcastle, Newcastle, New South Wales, Australia. [114]School of Electrical Engineering and Computer Science, University of Newcastle, Newcastle, New South Wales, Australia. [115]Division of Medical Genetics, Department of Biomedicine, University of Basel, Basel, Switzerland. [116]Department of Genetics, Harvard Medical School, Boston, MA, USA. [117]Section of Neonatal Screening and Hormones, Department of Clinical Biochemistry, Immunology and Genetics, Statens Serum Institut, Copenhagen, Denmark. [118]Department of Psychiatry, Fujita Health University School of Medicine, Toyoake, Japan. [119]Regional Centre for Clinical Research in Psychosis, Department of Psychiatry, Stavanger University Hospital, Stavanger, Norway. [120]Rheumatology Research Group, Vall d'Hebron Research Institute, Barcelona, Spain. [121]Centre for Medical Research, The University of Western Australia, Perth, Western Australia, Australia. [122]Department of Medical Genetics, Medical University, Sofia, Bulgaria. [123]Department of Psychology, University of Colorado Boulder, Boulder, CO, USA. [124]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada. [125]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada. [126]Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada. [127]Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia. [128]Latvian Biomedical Research and Study Centre, Riga, Latvia. [129]Department of Psychiatry and Zilkha Neurogenetics Institute, Keck School of Medicine at University of Southern California, Los Angeles, CA, USA. [130]Faculty of Medicine, Vilnius University, Vilnius, Lithuania. [131]Department of Biology and Medical Genetics, 2nd Faculty of Medicine and University Hospital Motol, Prague, Czech Republic. [132]Department of Child and Adolescent Psychiatry, Pierre and Marie Curie Faculty of Medicine, Paris, France. [133]Duke-NUS Graduate Medical School, Singapore, Singapore. [134]Department of Psychiatry, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [135]Centre for Genomic Sciences, The University of Hong Kong, Hong Kong, China. [136]Mental Health Centre and Psychiatric Laboratory, West China Hospital, Sichuan University, Chengdu, China. [137]Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA. [138]Department of Psychiatry, Columbia University, New York, NY, USA. [139]Department of Genetics and Pathology, International Hereditary Cancer Center, Pomeranian Medical University in Szczecin, Szczecin, Poland. [140]Department of Mental Health and Substance Abuse Services, National Institute for Health and Welfare, Helsinki, Finland. [141]Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [142]Department of Psychiatry, University of Bonn, Bonn, Germany. [143]Centre National de la Recherche Scientifique, Laboratoire de Génétique Moléculaire de la Neurotransmission et des Processus Neurodégénératifs, Hôpital de la Pitié Salpêtrière, Paris, France. [144]Department of Genomics Mathematics, University of Bonn, Bonn, Germany. [145]Research Unit, Sørlandet Hospital, Kristiansand, Norway. [146]Department of Psychiatry, National University of Ireland Galway, Galway, Ireland. [147]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK. [148]Division of Psychiatry, University of Edinburgh, Edinburgh, UK. [149]Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway. [150]Massachusetts Mental Health Center Public Psychiatry Division of the Beth Israel Deaconess Medical Center, Boston, MA, USA. [151]Estonian Genome Center, University of Tartu, Tartu, Estonia. [152]School of Psychology, University of Newcastle, Newcastle, New South Wales, Australia. [153]First Psychiatric Clinic, Medical University, Sofia, Bulgaria. [154]Department P, Aarhus University Hospital, Risskov, Denmark. [155]Department of Psychiatry, Royal College of Surgeons in Ireland, Dublin, Ireland. [156]King's College London, London, UK. [157]South Limburg Mental Health Research and Teaching Network, EURON, Maastricht University Medical Centre, Maastricht, The Netherlands. [158]Institute of Translational Medicine, University of Liverpool, Liverpool, UK. [159]Max Planck Institute of Psychiatry, Munich, Germany. [160]Munich Cluster for Systems Neurology (SyNergy), Munich, Germany. [161]Department of Psychiatry and Psychotherapy, Jena University Hospital, Jena, Germany. [162]Department of Psychiatry, Queensland Brain Institute and Queensland Centre for Mental Health Research, University of Queensland, Brisbane, Queensland, Australia. [163]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [164]Department of Psychiatry, Trinity College Dublin, Dublin, Ireland. [165]Eli Lilly, Lilly Corporate Center, Indianapolis, IN, USA. [166]Department of Clinical Sciences, Psychiatry, Umeå University, Umeå, Sweden. [167]DETECT Early Intervention Service for Psychosis, Blackrock, Ireland. [168]Centre for Public Health, Institute of Clinical Sciences, Queen's University Belfast, Belfast, UK. [169]Lawrence Berkeley National Laboratory, University of California at Berkeley, Berkeley, CA, USA. [170]Institute of Psychiatry, King's College London, London, UK. [171]Melbourne Neuropsychiatry Centre, University of Melbourne and Melbourne Health, Melbourne, Victoria, Australia. [172]Department of Psychiatry, University of Helsinki, Helsinki, Finland. [173]Public Health Genomics Unit, National Institute for Health and Welfare, Helsinki, Finland. [174]Medical Faculty, University of Belgrade, Belgrade, Serbia. [175]Department of Psychiatry, University of North Carolina, Chapel Hill, NC, USA. [176]Institute for Molecular Medicine Finland, FIMM, University of Helsinki, Helsinki, Finland. [177]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA. [178]Department of Psychiatry, University of Oxford, Oxford, UK. [179]Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA. [180]Institute for Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [181]PharmaTherapeutics Clinical Research, Pfizer Worldwide Research and Development, Cambridge, MA, USA. [182]Department of Psychiatry and Psychotherapy, University of Gottingen, Göttingen, Germany. [183]Psychiatry and Psychotherapy Clinic, University of Erlangen, Erlangen, Germany. [184]Hunter New England Health Service, Newcastle, New South Wales, Australia. [185]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. [186]University of Iceland, Landspitali, National University Hospital, Reykjavik, Iceland. [187]Department of Psychiatry and Drug Addiction, Tbilisi State Medical University (TSMU), Tbilisi, Georgia. [188]Research and Development, Bronx Veterans Affairs Medical Center, New York, NY, USA. [189]Wellcome Trust Centre for Human Genetics, Oxford, UK. [190]deCODE Genetics, Reykjavik, Iceland. [191]Department of Clinical Neurology, Medical University of Vienna, Vienna, Austria. [192]Lieber Institute for Brain Development, Baltimore, MD, USA. [193]Department of Medical Genetics, University Medical Centre Utrecht, Utrecht, The Netherlands. [194]Berkshire Healthcare NHS Foundation Trust, Bracknell, UK. [195]Section of Psychiatry, University of Verona, Verona, Italy. [196]Department of Psychiatry, University of Oulu, Oulu, Finland. [197]University Hospital of Oulu, Oulu, Finland. [198]Molecular and Cellular Therapeutics, Royal College of Surgeons in Ireland, Dublin, Ireland. [199]Health Research Board, Dublin, Ireland. [200]School of Psychiatry and Clinical Neurosciences, The University of Western Australia, Perth, Western Australia, Australia. [201]Computational Sciences CoE, Pfizer Worldwide Research and Development, Cambridge, MA, USA. [202]Human Genetics, Genome Institute of Singapore, A*STAR, Singapore, Singapore. [203]University College London, London, UK. [204]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [205]Institute of Neuroscience and Medicine (INM-1), Research Center Juelich, Juelich, Germany. [206]Department of Genetics, The Hebrew University of Jerusalem, Jerusalem, Israel. [207]Neuroscience Discovery and Translational Area, Pharma Research and Early Development, F. Hoffman-La Roche, Basel, Switzerland. [208]The Perkins Institute for Medical Research, The University of Western Australia, Perth, Western Australia, Australia. [209]Centre for Clinical Research in Neuropsychiatry, School of Psychiatry and Clinical Neurosciences, The University of Western Australia, Perth, Western Australia, Australia. [210]Virginia Institute for Psychiatric and Behavioral Genetics, Departments of Psychiatry and Human and Molecular Genetics, Virginia Commonwealth

University, Richmond, VA, USA. [211]The Feinstein Institute for Medical Research, Manhasset, NY, USA. [212]The Hofstra NS-LIJ School of Medicine, Hempstead, NY, USA. [213]The Zucker Hillside Hospital, Glen Oaks, NY, USA. [214]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore. [215]Queensland Centre for Mental Health Research, University of Queensland, Brisbane, Queensland, Australia. [216]Center for Human Genetic Research and Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [217]Department of Child and Adolescent Psychiatry, Erasmus University Medical Centre, Rotterdam, The Netherlands. [218]Department of Complex Trait Genetics, Neuroscience Campus Amsterdam, VU University Medical Center Amsterdam, Amsterdam, The Netherlands. [219]Department of Functional Genomics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University, Amsterdam, The Netherlands. [220]University of Aberdeen, Institute of Medical Sciences, Aberdeen, UK. [221]Departments of Psychiatry, Neurology, Neuroscience and Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA. [222]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. [223]Fundació de Docència i Recerca Mútua de Terrassa, Universitat de Barcelona, Barcelona, Spain. [224]Child and Adolescent Psychiatry, University of Technology Dresden, Dresden, Germany. [225]Section for Experimental Psychopathology, General Psychiatry, Heidelberg, Germany. [226]Institute of Cognitive Neuroscience, University College London, London, UK. [227]Mental Health Sciences Unit, University College London, London, UK. [228]Centro Investigación Biomédica en Red Salud Mental, Madrid, Spain. [229]University Hospital Marqués de Valdecilla, Instituto de Formación e Investigación Marqués de Valdecilla, University of Cantabria, Santander, Spain. [230]Centre for Clinical Research in Neuropsychiatry, The University of Western Australia, Perth, Western Australia, Australia. [231]Western Australian Institute for Medical Research, The University of Western Australia, Perth, Western Australia, Australia. [232]Department of Psychiatry, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. [233]Department of Psychiatry, University Medical Center Groningen, University of Groningen, The Netherlands. [234]Department of Statistics, University of Oxford, Oxford, UK. [235]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK. [236]Cambridge Institute for Medical Research, University of Cambridge School of Clinical Medicine, Cambridge, UK. [237]Telethon Institute for Child Health Research, Centre for Child Health Research, University of Western Australia, Subiaco, Western Australia, Australia. [238]Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Brisbane, Queensland, Australia. [239]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. [240]Department of Epidemiology and Public Health, University College London, London, UK. [241]Molecular and Physiological Sciences, The Wellcome Trust, London, UK. [242]Peninsula School of Medicine and Dentistry, Plymouth University, Plymouth, UK. [243]Clinical Neurosciences, St George's University of London, London, UK. [244]Department of Medical and Molecular Genetics, School of Medicine, Guy's Hospital, King's College London, London, UK. [245]Biomedical Research Centre, Ninewells Hospital and Medical School, Dundee, UK. [246]Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. [247]Institute of Ophthalmology, University College London, London, UK. [248]National Institute for Health Research, Biomedical Research Centre at Moorfields Eye Hospital, National Health Service Foundation Trust, London, UK. [249]Department of Molecular Neuroscience, Institute of Neurology, London, UK. [250]Oxford Centre for Diabetes, Endocrinology and Metabolism, Churchill Hospital, Oxford, UK.

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Creation of a C4 reference panel from WGS data

We constructed a reference panel for imputation of C4 structural haplotypes using WGS data for 1,265 individuals from the Genomic Psychiatry Cohort[26]. The reference panel included individuals of diverse ancestry, including 765 Europeans, 250 African Americans and 250 people of reported Latino ancestry.

We estimated the diploid C4 copy number, and estimated separately the diploid copy number of the contained human endogenous retrovirus (HERV) sequence, using Genome STRiP[44]. In brief, Genome STRiP carefully calibrates measurements of read depth across specific genomic segments of interest by estimating and normalizing away sample-specific technical effects such as the effect of GC content on read depth (estimated from the genome-wide data). To measure total C4 gene copy number, we analysed the segments 6:31948358–31981050 and 6:31981096–32013904 (hg19), masking the intronic HERV segments that distinguish short (S) from long (L) C4 gene isotypes. To measure copy number of the HERV sequence, we analysed segments 6:31952461–31958829 and 6:31985199–31991567 (hg19). Across the 1,265 individuals, the resultant locus-specific copy-number estimates exhibited a strongly multi-modal distribution (Extended Data Fig. 1a) from which individuals' total C4 copy numbers could be readily inferred.

We then estimated the numbers of *C4A* and *C4B* genes in each individual genome. To do this, we extracted reads mapping to the paralogous sequence variants that distinguish *C4A* from *C4B* (hg19 coordinates 6:31963859–31963876 and 6:31996597–31996614) in each individual, combining reads across the two sites. We included only reads that aligned to one of these segments in its entirety. We then counted the number of reads matching the canonical active site sequences for *C4A* (CCC TGT CCA GTG TTA GAC) and *C4B* (CTC TCT CCA GTG ATA CAT). We combined these counts with the likelihood estimates of diploid C4 copy number (from Genome STRiP) to determine the maximum likelihood combination of *C4A* and *C4B* in each individual (Extended Data Fig. 1b). We estimated the genotype quality of the *C4A* and *C4B* estimate from the likelihood ratio between the most likely and second most likely combinations.

To phase the C4 copy number measurements into haplotypes, we first used the GenerateHaploidCNVGenotypes utility in Genome STRiP to estimate haplotype-specific copy-number likelihoods for C4 (total C4 gene copy number), *C4A*, *C4B* and HERV using the diploid likelihoods from the prior step as input. Default parameters for GenerateHaploidCNVGenotypes were used, plus -genotypeLikelihoodThreshold 0.0001. The output was then processed by the GenerateCNVHaplotypes utility in Genome STRiP to combine the multiple estimates into likelihood estimates for a set of unified structural alleles. GenerateCNVHaplotypes was run with default parameters, plus -defaultLogLikelihood -50, -unknownHaplotypeLikelihood -50, and -sampleHaplotypePriorLikelihood 2.0. The resultant VCF output was phased using Beagle 4.1 (beagle_4.1_27Jul16.86a) in two steps: first, performing genotype refinement from the genotype likelihoods using the Beagle gtgl = and maxlr = 1000000 parameters, and then running Beagle again on the output file using gt = to complete the phasing.

Our previous work suggested that several C4 structures segregate on multiple haplotypes, and probably arose by recurrent mutation on different haplotype backgrounds[7]. The GenerateCNVHaplotypes utility requires as input an enumerated set of structural alleles to assign to the samples in the reference cohort, including any structurally equivalent alleles, with distinct labels to mark them as independent, plus a list of samples to assign (with high likelihood) to specific labelled input alleles to disambiguate among these recurrent alleles. The selection of the set of structural alleles to be modelled, along with the labelling strategy, is important to our methodology and the performance of the reference panel. In the reference panel, each input allele represents a specific copy number structure and optionally includes a label that differentiates the allele from other independent alleles with equivalent structure. We use the notation <H_n_n_n_n_L> to identify each allele, where the four integers following the H are, respectively, the (redundant) haploid count of the total number of C4 copies, *C4A* copies, *C4B* copies and HERV copies on the haplotype. For example, <H_2_1_1_1> was used to represent the 'AL-BS' haplotype. The optional final label L is used to distinguish potentially recurrent haplotypes with otherwise equivalent structures (under the model) that should be treated as independent alleles for phasing and imputation.

To build the reference panel, we experimentally evaluated a large number of potential sets of structural alleles and methods for assigning labels to potentially recurrent alleles. For each evaluation, we built a reference panel using the 1,265 reference samples, and then evaluated the performance of the panel via cross-validation, leaving out 10 different samples in each trial (5 samples in the last trial) and imputing the missing samples from the remaining samples in the panel. The imputed results for all 1,265 samples were then compared to the original diploid copy number estimates to evaluate the performance of each candidate reference panel (Extended Data Table 1).

Using this procedure, we selected a final panel for downstream analysis that used a set of 29 structural alleles representing 16 distinct allelic structures (as listed in the reference panel VCF file). Each allele contained from one to three copies of C4. Three allelic structures (AL-BS, AL-BL and AL-AL) were represented as a set of independently labelled alleles with 9, 3 and 4 labels, respectively.

To identify the number of labels to use on the different alleles and the samples to 'seed' the alleles, we generated spider plots of the C4 locus based on initial phasing experiments run without labelled alleles, and then clustered the resulting haplotypes in two dimensions based on the *y*-coordinate distance between the haplotypes on the left and right sides of the spider plot. Clustering was based on visualizing the clusters (Extended Data Fig. 1c) and then manually choosing both the number of clusters (labels) to assign and a set of confidently assigned haplotypes to use to seed the clusters in GenerateCNVHaplotypes. This procedure was iterated multiple times using cross-validation, as described above, to evaluate the imputation performance of each candidate labelling strategy.

Within the dataset used to build the reference panel, there is evidence for individuals carrying seven or more diploid copies of C4, which implies the existence of (rare) alleles with four or more copies of C4. In our experiments, attempting to add additional haplotypes to model these rare four-copy alleles reduced overall imputation performance. Consequently, we conducted all downstream analyses using a reference panel that models only alleles with up to three copies of C4. In the future, larger reference panels might benefit from modelling these rare four-copy alleles.

The reference panel will be available in dbGaP (accession number pending) with broad permission for research use.

Genetic data for SLE

For analysis of SLE, collection and genotyping of the European-ancestry cohort (6,748 cases, 11,516 controls, genotyped by ImmunoChip) as previously described[3]. Collection and genotyping of the African American cohort (1,494 cases, 5,908 controls, genotyped by OmniExpress) as previously described[5].

Genetic data for Sjögren's syndrome

For analysis of Sjögren's syndrome, collection and genotyping of the European-ancestry cohort (673 cases, 1,153 controls, genotyped by Omni2.5) as previously described[32] and available in dbGaP under study accession number phs000672.v1.p1.

Genetic data for schizophrenia

The schizophrenia analysis made use of genotype data from 40 cohorts of European ancestry (28,799 cases, 35,986 controls) made available by the Psychiatric Genetics Consortium (PGC) as previously described[43]. Genotyping chips used for each cohort are listed in supplementary table 3 of that study.

Imputation of C4 alleles

The reference haplotypes described above were used to extend the SLE, Sjögren's syndrome or schizophrenia cohort SNP genotypes by imputation. SNP data in VCF format were used as input for Beagle v.4.1[45,46] for imputation of C4 as a multi-allelic variant. Within the Beagle pipeline, the reference panel was first converted to bref format. From the cohort SNP genotypes, we used only those SNPs from the MHC region (chr6:24–34 Mb on hg19) that were also in the haplotype reference panel. We used the conform-gt tool to perform strand-flipping and filtering of specific SNPs for which strand remained ambiguous. Beagle was run using default parameters with two key exceptions: we used the GRCh37 PLINK recombination map, and we set the output to include genotype probability (that is, GP field in VCF) for correct downstream probabilistic estimation of *C4A* and *C4B* joint dosages.

Imputation of HLA alleles

For HLA allele imputation, sample genotypes were used as input for the R package HIBAG[47]. For both European ancestry and African American cohorts, publicly available multi-ethnic reference panels generated for the most appropriate genotyping chip (that is, Immunochip for European ancestry SLE cohort, Omni 2.5 for the European ancestry Sjögren's syndrome cohort, and OmniExpress for African American SLE cohort) were used[48]. Default parameters were used for all settings. All class I and class II HLA genes were imputed. Output haplotype posterior probabilities were summed per allele to yield diploid dosages for each individual.

Associating single and joint C4 structural allele dosages to SLE and Sjögren's syndrome in European ancestry individuals

The analysis described above yields dosage estimates for each of the common C4 structural haplotypes (for example, AL-BS or AL-AL) for each genome in each cohort. In addition to performing association analysis on these structures (Fig. 1b), we also performed association analysis on the dosages of each underlying C4 gene isotype (that is, *C4A*, *C4B*, *C4L* and *C4S*). These dosages were computed from the allelic dosage (DS) field of the imputation output VCF simply by multiplying the dosage of a C4 structural haplotype by the number of copies of each C4 isotype that haplotype contains (for example, AL-BL contains one *C4A* gene and one *C4B* gene).

C4 isotype dosages were then tested for disease association by logistic regression, with the inclusion of four available ancestry covariates derived from genome-wide principal component analysis (PCA) as additional independent variables, $PC_{c,}$

$$\text{logit}(\theta) = \beta_0 + \beta_1 \text{C4} + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (1)$$

where $\theta = \text{E}[\text{SLE}|\mathbf{X}]$, C4 is dosage of one of the isotypes per individual, β_0 is the fit intercept, other β values associated with each independent variable are best fit coefficients across the cohort, and ε is residual error. For Sjögren's syndrome, the model instead included two available multiethnic ancestry covariates from dbGaP that correlated strongly with European-specific ancestry covariates (specifically, PC5 and PC7) and smoking status as independent variables. Coefficients for relative weighting of *C4A* and *C4B* dosages (C4A and C4B) were obtained from a joint logistic regression,

$$\text{logit}(\theta) = \beta_0 + \beta_1 \text{C4A} + \beta_2 \text{C4B} + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (2)$$

where terms are as in equation (1) except both *C4A* and *C4B* isotype dosages are included.

The values per individual of $\beta_1\text{C4A} + \beta_2\text{C4B}$ were used as a combined C4 risk term for estimating both association strength (Extended Data Fig. 3a, b) as well as evaluating the relationship between the strength of nearby variants' association with SLE or Sjögren's syndrome and linkage with C4 variation (Extended Data Fig. 4a–c).

Joint dosages of *C4A* and *C4B* for each individual in the same cohort were estimated by summing across their genotype probabilities of paired structural alleles that encode for the same diploid copy numbers of both *C4A* and *C4B* (Extended Data Fig. 2a, b). For each individual or genome, this yields a joint dosage distribution of *C4A* and *C4B* gene copy number, reflecting any possible imputed haplotype-level dosages with non-zero probability. Joint dosages for *C4A* and *C4B* diploid copy numbers were tested for association with SLE in a joint model with the same ancestry covariates (Fig. 1a),

$$\text{logit}(\theta) = \beta_0 + \sum_{i,j} \beta_{i,j} P(\text{C4A} = i, \text{C4B} = j) + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (3)$$

where terms are as in equation (1) except $P(\text{C4A} = i, \text{C4B} = j)$ which represents the probability that an individual has i integer copies of *C4A* and j integer copies of *C4B*.

Calculation of composite C4 risk for SLE

SLE risk was strongly associated with *C4A* and *C4B* copy numbers (Fig. 1a) in an initial, simple model in which their contributions were treated as linear and independent. In specific subsequent analyses (for example, to map C4-independent effects), to account for the possibility of nonlinear or interacting contributions, a composite C4 risk score was derived by taking the weighted sum of joint *C4A* and *C4B* dosages multiplied by the corresponding effect sizes from the aforementioned model of the joint *C4A* and *C4B* diploid copy numbers. The weights for calculating this composite C4 risk term were computed from the data from the European ancestry cohort, and then applied unchanged to analysis of the African American cohort.

Associations of variants across the MHC region to SLE and Sjögren's syndrome

Genotypes for non-array SNPs were imputed with IMPUTE2 using the 1,000 Genomes reference panel; separate analyses were performed for the European-ancestry and African American cohorts. Unless otherwise stated, all subsequent SLE analyses were performed identically for both European ancestry and African American cohorts. Dosage of each variant, v_i, was tested for association with SLE or Sjögren's syndrome in a logistic regression including available ancestry covariates (and smoking status for Sjögren's syndrome) first alone (Extended Data Fig. 3a, b),

$$\text{logit}(\theta) = \beta_0 + \beta_1 v_i + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (4)$$

then with C4 composite risk (Extended Data Fig. 3c),

$$\text{logit}(\theta) = \beta_0 + \beta_1 v_i + \beta_1 \text{C4} + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (5)$$

where other terms are as in equation (1). For Sjögren's syndrome, the simpler weighted (2.3)C4A + C4B model was used instead of composite risk term, as the cohort's size gave poor precision to estimates of risk for many joint (C4A, C4B) copy numbers (Extended Data Fig. 3d). The Pearson correlation between the C4 composite risk term and each other variant was computed and squared (r^2) to yield a measure of LD between C4 composite risk and that variant in that cohort.

Association analyses for specific C4 structural alleles

The C4 structural haplotypes were tested for association with disease (Figs. 1b, 2a) in a joint logistic regression that included (1) terms for dosages of the five most common C4 structural haplotypes (AL-BS, AL-BL,

AL-AL, BS and AL), (2) (for SLE and Sjögren's syndrome) rs2105898 genotype, and (3) ancestry covariates and (for Sjögren's syndrome) smoking status,

$$\text{logit}(\theta) = \beta_0 + \beta_1 \text{BS} + \beta_2 \text{AL} + \beta_3 \text{ALBS} + \beta_4 \text{ALBL} + \beta_5 \text{ALAL} + \beta_6 \text{rs2105898} + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (6)$$

where other terms are as in equation (1). Several of these common C4 structural alleles arose multiple times on distinct haplotypes; we term the set of haplotypes in which such a common allele appeared as haplogroups. The haplogroups can be further tested in a logistic regression model in which the structural allele appearing in all member haplotypes is instead encoded as dosages for each of the SNP haplotypes in which it appears. These association analyses (Figs. 1b, 2a) were performed as in equation (6), with structural allele dosages for ALBS, ALBL and ALAL replaced by multiple terms for each distinct haplotype.

To delineate the relationship between C4-BS and *DRB1**03:01 alleles—which are highly linked in European ancestry haplotypes—allelic dosages per individual in the African American SLE cohort were rounded to yield the most likely integer dosage for each. Although genotype dosages for each are reported by BEAGLE and HIBAG respectively, probabilities per haplotype are not linked and multiplying possible diploid dosages could yield incorrect non-zero joint dosages. Joint genotypes were tested as individual terms in a logistic regression model (Fig. 2b),

$$\text{logit}(\theta) = \beta_0 + \sum_{i,j} \beta_{i,j} P(\text{C4–BS} = i, \text{DRB1*03 : 01} = j) + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (7)$$

where terms are as in equation (1) except P(C4-BS = i,DRB1*03:01 = j) which represents the probability that an individual has i haplotypes with C4-BS allele and j haplotypes with *DRB1**03:01 allele.

Sex-stratified associations of C4 structural alleles and other variants with SLE, Sjögren's syndrome and schizophrenia

Determination of an effect from sex on the contribution of overall C4 variation to risk for each disorder was done by including an interaction term between sex and C4; that is, (2.3)C4A + C4B for SLE and Sjögren's syndrome and estimated C4A expression for schizophrenia:

$$\text{logit}(\theta) = \beta_0 + \beta_2 \text{C4} + \beta_3 I_{\text{sex}} + \beta_4 I_{\text{sex}} \text{C4} + \sum_c \beta_c \text{PC}_c + \varepsilon \quad (8)$$

where terms are as in equation (1) except the term C4 = (2.3)C4A + C4B and I_{sex} which is an indicator variable for whether an individual is male.

Each variant in the MHC region was tested for association with among European ancestry cases and cohorts in a logistic regression as in equations (4)–(6) using only male cases and controls, and then separately using only female cases and controls (Extended Data Fig. 6a–c). Likewise, allelic series analyses were performed as in equation (7), but in separate models for men and women (Fig. 3a, b).

To assess the relationship between sex bias in the risk associated with a variant and linkage to C4 composite risk (as non-negative r^2), male and female log-odds were multiplied by the sign of the Pearson correlation between that variant and C4 composite risk before taking the difference.

Analyses of CSF

CSF from healthy individuals was obtained from two research panels. The first panel, consisting of 533 donors (327 male, 126 female) from hospitals around Utrecht, Netherlands, was described previously[49,50]. The donors were generally healthy research participants undergoing spinal anaesthesia for minor elective surgery. The same donors were previously genotyped using the Illumina Omni SNP array. To estimate C4 copy numbers, we used SNPs from the MHC region (chr6:24-34 Mb on hg19) as input for C4 allele imputation with Beagle, as described above in 'Imputation of C4 alleles'.

The second CSF panel sampled specimens from 56 donors (14 male, 42 female) from Brigham and Women's Hospital (BWH) under a protocol approved by the institutional review board at BWH (IRB protocol ID no. 1999P010911) with informed consent. These samples were originally obtained to exclude the possibility of infection, and clinical analyses had revealed no evidence of infection. Donors ranged from 18 to 64 years of age. Blood samples from the same individuals were used for extraction of genomic DNA, and C4 gene copy number was measured by droplet digital PCR (ddPCR) as previously described[7]. Samples were excluded from measurements if they lacked C4 genotypes, sex information, or contained visible blood contamination.

C4 measurements were performed by sandwich ELISA of 1:400 dilutions of the original CSF sample using goat anti-sera against human C4 as the capture antibody (Quidel, A305, used at 1:1,000 dilution), FITC-conjugated polyclonal rabbit anti-human C4c as the detection antibody (Dako, F016902-2, used at 1:3,000 dilution), and alkaline phosphatase–conjugated polyclonal goat anti-rabbit IgG as the secondary antibody (Abcam, ab97048, used at 1:5,000 dilution). C3 measurements were performed using the human complement C3 ELISA kit (Abcam, ab108823).

Because C4 gene copy number had a large and proportional effect on C4 protein concentration in these CSF samples (Extended Data Fig. 7a), we corrected for C4 gene copy number in our analysis of relationship between sex and C4 protein concentration, by normalizing the ratio of C4 protein (in CSF) to C4 gene copies (in genome). Therefore, these analyses included only samples for which DNA was available or C4 was successfully imputed. In total, 495 (332 male, 163 female) C4 and 304 (179 male, 125 female) C3 concentrations were obtained across both cohorts. log concentrations of C3 (in ng ml^{-1}) and C4 (in ng ml^{-1}, per C4 gene copy number) protein were then used separately in linear regression models to estimate a sex-unbiased cohort-specific offset for each protein,

$$\log_{10}(\text{C3 or C4 concentration}) = \beta_0 + \beta_1 I_{\text{sex}} + \beta_2 I_{\text{cohort}} + \varepsilon \quad (9)$$

to be applied to all concentrations for that protein, where I_{sex} is an indicator variable for whether an individual is male, I_{cohort} is an indicator variable for whether an individual was in the second cohort, β_0 is the fit intercept, other β associated with each independent variable are best fit coefficients across the cohort, and ε is residual error. Estimation of average measurements by age for each sex was done by LOESS (Fig. 3c, d). To evaluate the significance of sex effects, we used these cohort-corrected concentrations estimates and analysed them with the non-parametric unsigned Mann–Whitney rank-sum test comparing concentration distributions for males and females.

Analyses of blood plasma

Blood plasma was collected and immunoturbidimetric measurements of C3 and C4 protein in 1,844 individuals (182 men, 1662 women) by Sjögren's International Collaborative Clinical Alliance (SICCA) from individuals with and without Sjögren's syndrome as previously described[51]. C4 copy numbers for these individuals were previously imputed for use in logistic regression of Sjögren's syndrome risk. As C4 copy number has an effect on measured C4 protein similar to CSF (Extended Data Fig. 7b), we normalized C4 levels to them in all following analyses. Estimation of average measurements by age for each sex was done by local polynomial regression smoothing (LOESS) on log-concentrations of C3 (mg dl^{-1}) and C4 (mg dl^{-1}, per C4 gene copy number) protein (Extended Data Fig. 7c, d). To evaluate the significance of sex bias within age ranges displaying the greatest difference (informed by LOESS), we analysed individuals in these bins with the

non-parametric unsigned Mann–Whitney rank-sum test comparing concentration distributions for males and females.

Difference in C4 protein levels between individual with and without Sjögren's syndrome was done by performing a non-parametric unsigned Mann–Whitney rank-sum test on C4 protein levels with and without normalization to C4 genomic copy number (Extended Data Fig. 7e, f).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Individual genotype data for Sjögren's syndrome cases and controls and individual plasma concentrations for C4 and C3 are available in dbGaP under accession number phs000672.v1.p1. Individual genotype data for schizophrenia cases and controls are available by application to the Psychiatric Genomics Consortium (PGC). Questions regarding individual genotype data for SLE cases and controls of European and/or African American ancestry can be directed to T.J.V. Data resources are available on the McCarroll lab website at http://mccarrolllab.org/resources/resources-for-c4/. We have deposited the haplotype reference panel we created for C4 imputation in dbGaP under accession number phs001992.v1.p1. Genotype and protein concentration data for CSF samples are available upon request.

Code availability

Software scripts and instructions for imputing C4 alleles into SNP datasets are available on the McCarroll laboratory website at http://mccarrolllab.org/resources/resources-for-c4/.

44. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
45. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
46. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
47. Zheng, X. et al. HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).
48. Zheng, X. Imputation-based HLA typing with SNPs in GWAS studies. *Methods Mol. Biol.* **1802**, 163–176 (2018).
49. Luykx, J. J. et al. A common variant in ERBB4 regulates GABA concentrations in human cerebrospinal fluid. *Neuropsychopharmacology* **37**, 2088–2092 (2012).
50. Albersen, M. et al. Vitamin B-6 vitamers in human plasma and cerebrospinal fluid. *Am. J. Clin. Nutr.* **100**, 587–592 (2014).
51. Malladi, A. S. et al. Primary Sjögren's syndrome as a systemic disease: a study of participants enrolled in an international Sjögren's syndrome registry. *Arthritis Care Res. (Hoboken)* **64**, 911–918 (2012).
52. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
53. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

Acknowledgements This work was supported by the National Human Genome Research Institute (HG006855), the National Institute of Mental Health (MH112491, MH105641, MH105653), the Stanley Center for Psychiatric Research, and the National Institute for Health Research Biomedical Research Centre (NIHR BRC) at Guy's and St Thomas' NHS Foundation and King's College London. We thank C. Usher and C. Patil for contributions to the figures and manuscript text, and M. Florio for suggestions regarding figure display.

Author contributions N.K., A.S., T.J.V. and S.A.M. conceived the genetic studies. M.T.P., C.N.P. and M.B. collected and contributed WGS data for the Genomic Psychiatry Cohort. R.E.H. and C.W.W. genotyped C4 structural variation in the Genomic Psychiatry Cohort and optimized variant selection for use as a reference panel in the imputation of C4 variation into lupus and schizophrenia cohorts (Extended Data Fig. 1). T.J.V., R.R.G., L.A.C., C.D.L., R.P.K., J.B.H., K.M.K., D.L.M. and P.T. contributed genotype data and imputation of non-C4 variation for analysis of SLE cohorts. K.E.T. and L.A.C. contributed genotype and phenotype data along with imputation of non-C4 variation for analysis of the Sjögren's syndrome cohort. Investigators in the Schizophrenia Working Group of the Psychiatric Genomics Consortium collected and phenotyped cohorts and contributed genotype data for analysis of schizophrenia cohorts. N.K. did the imputation and association analysis (Figs. 1, 2, 3a, b, Extended Data Figs. 2–6). T.J.V., R.R.G. and D.L.M. provided valuable advice on the analysis and interpretation of SLE-association results. R.A.O. and L.M.O.L. collected and provided CSF samples composing the group from Utrecht, Netherlands. C.E.S. collected and provided CSF samples composing the Brigham & Women's Hospital group. H.d.R and K.T. performed the C4 and C3 immunoassay experiments on CSF samples (Fig. 3c, d, Extended Data Fig. 7a). N.K. did the analysis of plasma C4 and C3 concentrations (Extended Data Fig. 7b–f). S.A.M and N.K. wrote the manuscript with contributions from all authors. Management Committee of Wellcome Trust Case–Control Consortium 2: P.D., I.B., J.M.B., E.B., M.A.B., J.P.C., A.C., P.D., A.D., J.J., H.S.M., C.G.M., C.N.A.P., R.P., A.R., S.J.S., R.C.T., A.C.V. and N.W.W; Data and Analysis Group of Wellcome Trust Case–Control Consortium 2: C.C.A.S., G.B., C.B., P.D., C.F., E.G., G.H., R.P., M.P., A.S., Z.S., D.V.; DNA, Genotyping, Data QC, and Informatics of Wellcome Trust Case–Control Consortium 2: C.L., I.B., H.B., S.J.B., P.D., S.D., S.E., M.G., E.G., R.G., N.H., S.E.H., A.J., J.L., O.T.M., S.C.P., R.R., M.R., A.T.-G., M.W., P.W., P.W., S.W.; Publications Committee of Wellcome Trust Case–Control Consortium 2: C.G.M., J.M.B., M.A.B., A.C., M.I..M. and C.C.A.S.

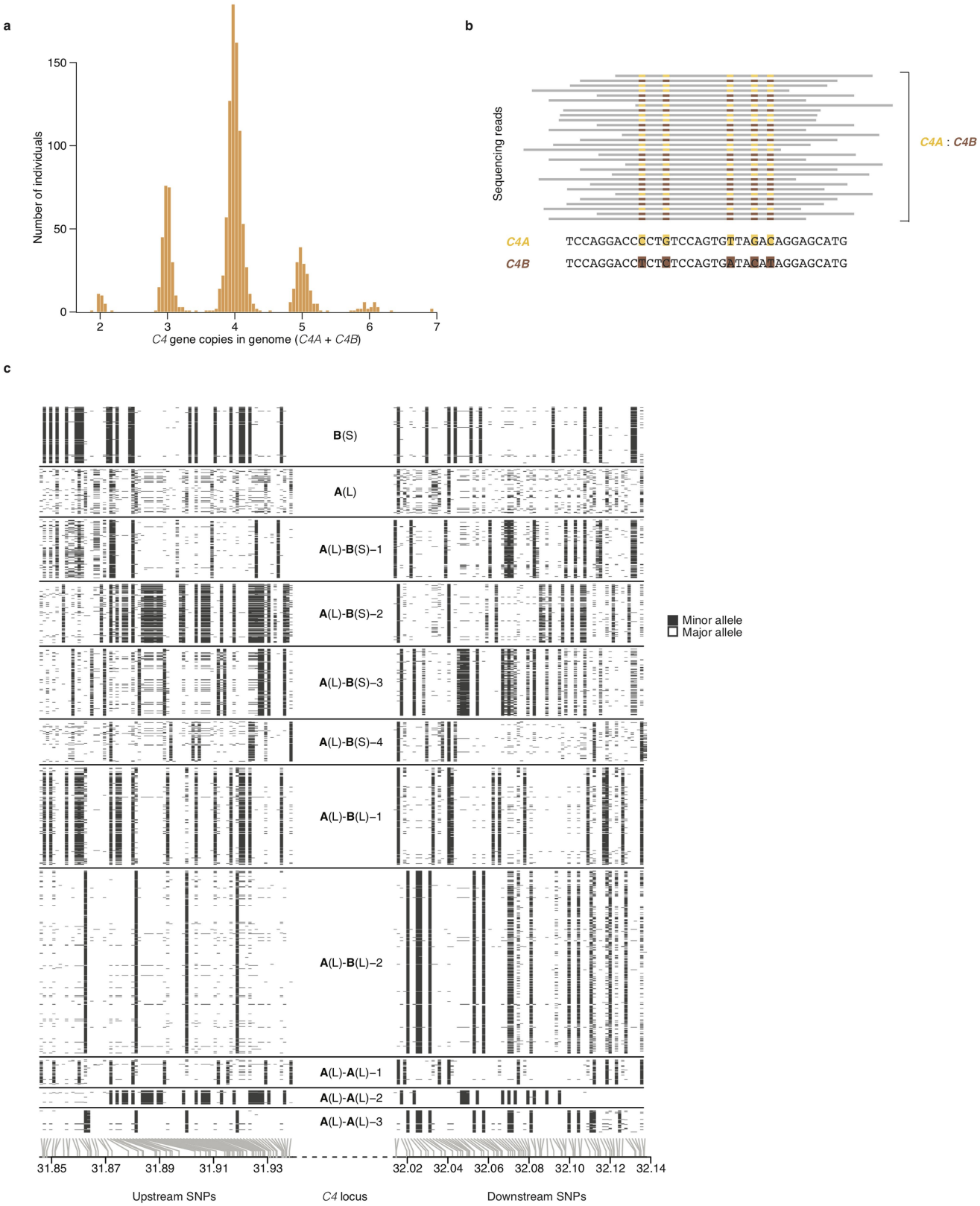
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2277-x.

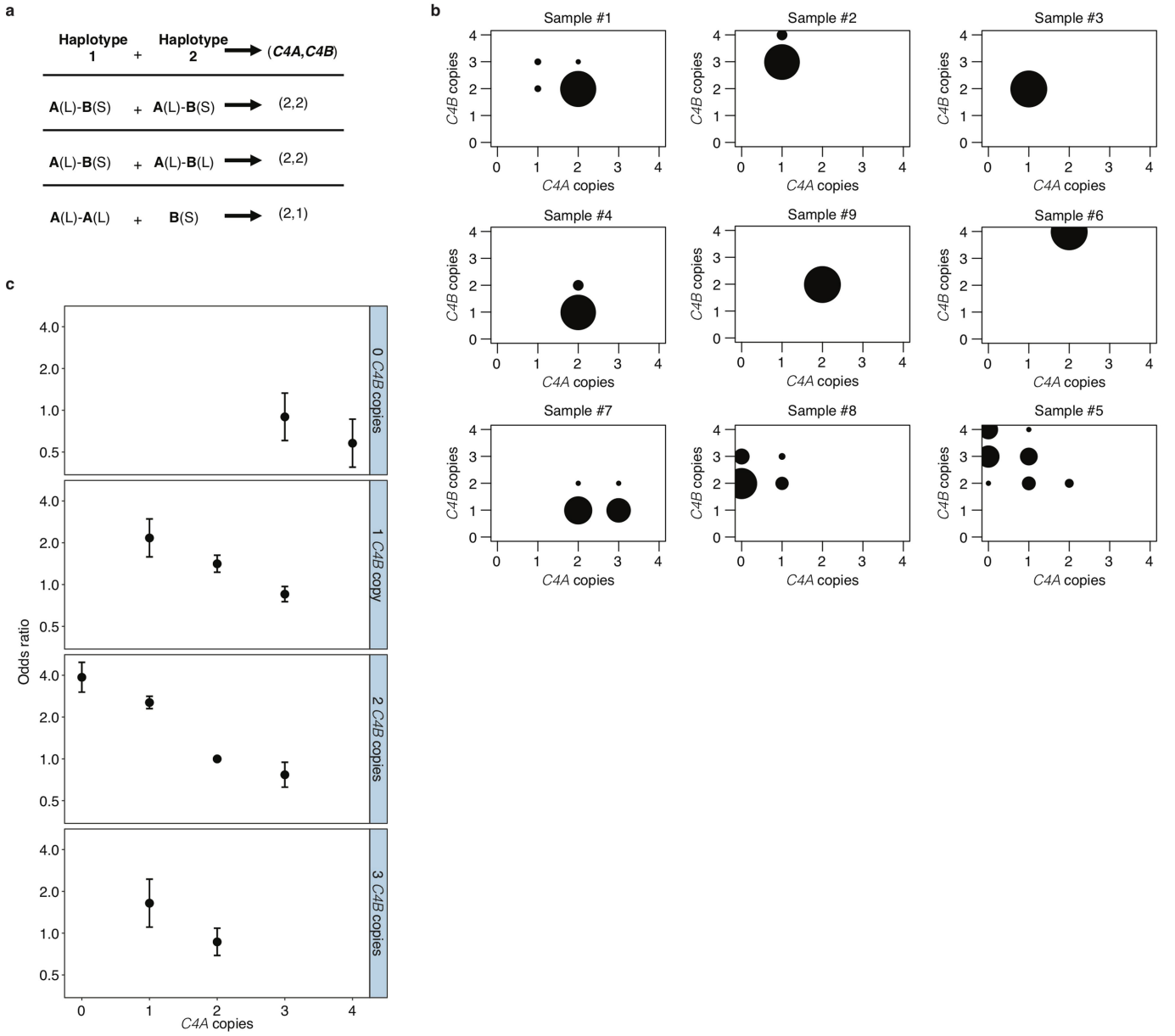
Correspondence and requests for materials should be addressed to N.K., T.J.V. or S.A.M.

Peer review information *Nature* thanks John Armour and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

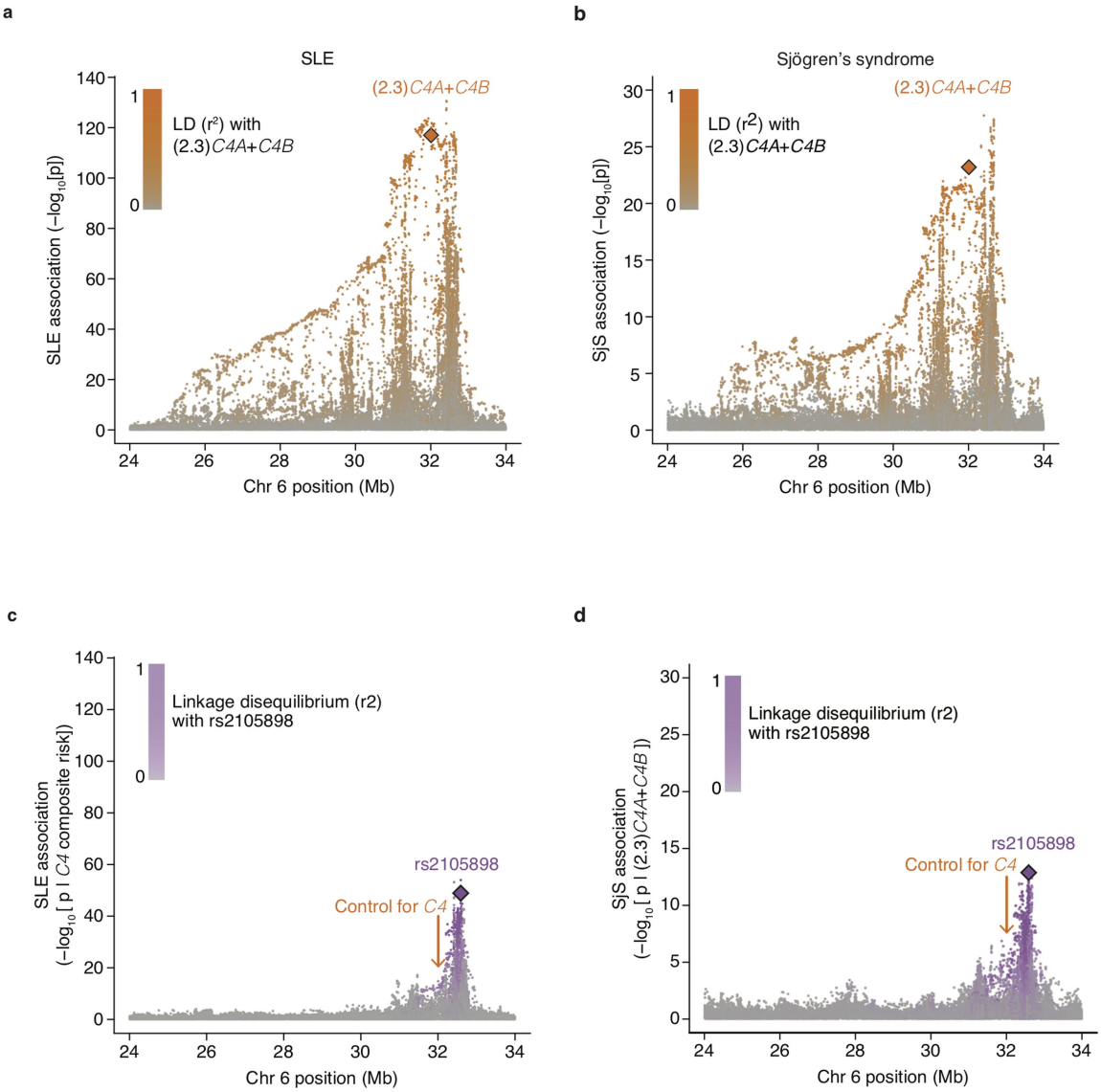


Extended Data Fig. 1 | See next page for caption.

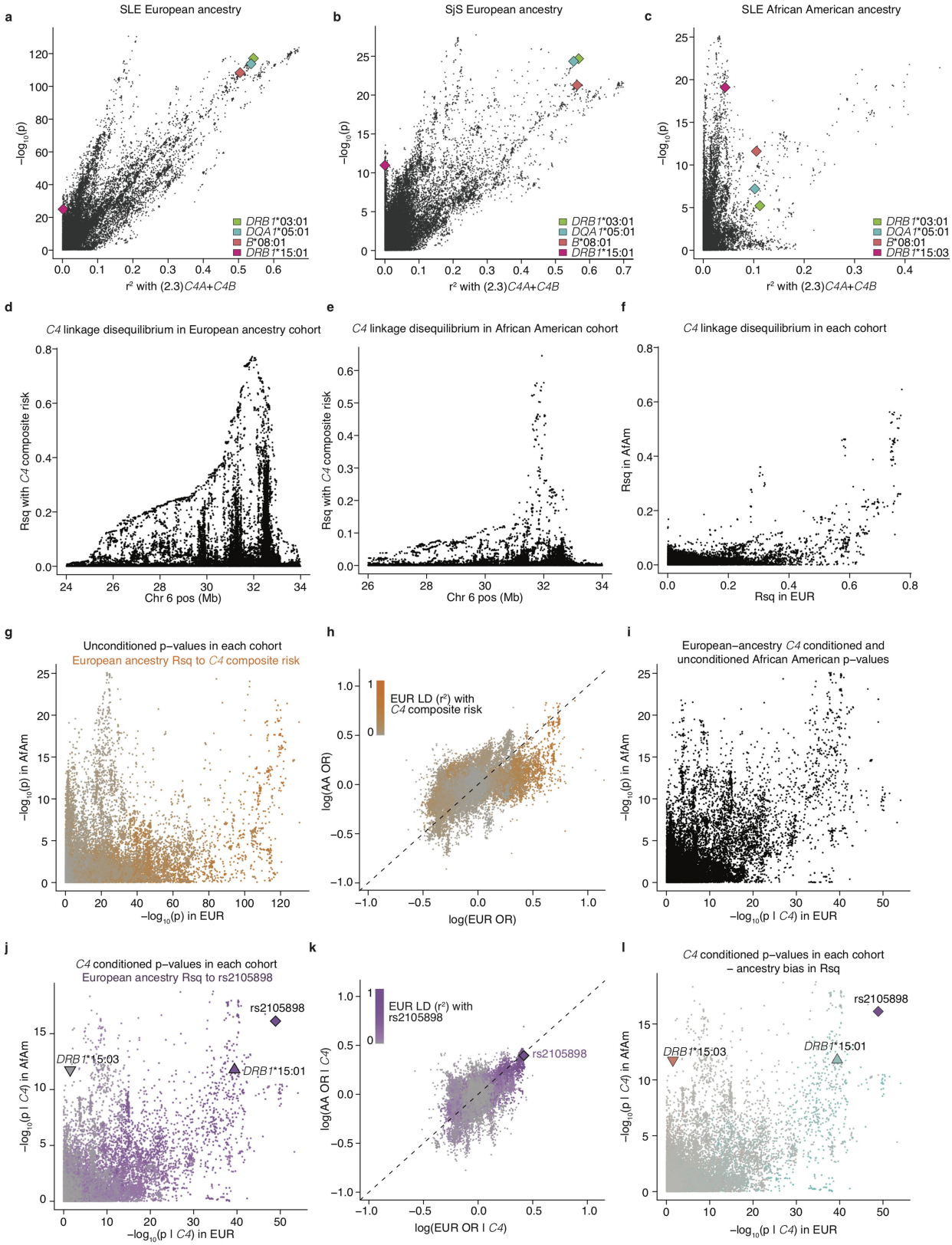
Extended Data Fig. 1 | A panel of 2,530 reference haplotypes (created from WGS data) containing C4 alleles and SNPs across the MHC genomic region enables imputation of C4 alleles into SNP data. a, Distributions (across 1,265 individuals) of total C4 gene copy number (*C4A* + *C4B*), as measured from read depth of coverage across the C4 locus, in WGS data. **b**, The relative numbers of reads that overlap sequences specific to *C4A* or *C4B* (together with the total C4 gene copy number as in **a**) are used to infer the underlying copy numbers of the *C4A* and *C4B* genes. For example, in an individual with four C4 genes, the presence of equal numbers of reads specific to *C4A* or *C4B* suggests the presence of two copies each of *C4A* and *C4B*. Precise statistical approaches (including inference of probabilistic dosages) and further approaches for phasing C4 allelic states with nearby SNPs to create reference haplotypes, are described in Methods. **c**, The SNP haplotypes flanking each C4 allele are shown as rows (SNPs as columns), with white and black representing the major and minor allele of each SNP. Grey lines at the bottom indicate the physical location of each SNP along chromosome 6. The differences among the haplotypes are most pronounced closest to C4 (towards the centre of the plot), as historical recombination events in the flanking megabases will have caused the haplotypes to be less consistently distinct at greater genomic distances from C4. The patterns indicate that many combinations of *C4A* and *C4B* gene copy numbers have arisen recurrently on more than one SNP haplotype, a relationship that can be used in association analyses (Fig. 1b).



Extended Data Fig. 2 | Aggregation of joint *C4A* and *C4B* genotype probabilities per individual across imputed C4 structural alleles for estimation of SLE risk for each combination. a, An individual's joint *C4A* and *C4B* gene copy number can be calculated by summing the *C4A* and *C4B* gene contents for each possible pair of two inherited alleles. Many pairings of possible inherited alleles result in the same joint *C4A* and *C4B* gene copy number. **b**, Each individual's *C4A* and *C4B* gene copy number was imputed from their SNP data, using the reference haplotypes summarized in Extended Data Fig. 1c. For more than 95% of individuals (exemplified by samples 1–6 in the figure), this inference can be made with >90% certainty or confidence (the areas of the circles represent the posterior probability distribution over possible *C4A/C4B* gene copy numbers). For the remaining individuals (exemplified by samples 7–9 in the figure), greater statistical uncertainty persists about C4 genotype. To account for this uncertainty, in downstream association analysis, all C4 genotype assignments are handled as probabilistic gene dosages—analogous to the genotype dosages that are routinely used in large-scale genetic association studies that use imputation. **c**, Odds ratios and 95% confidence intervals underlying each of the C4-genotype risk estimates in Fig. 1a presented as a series of panels for each observed copy number of *C4B*, with increasing copy number of *C4A* for that *C4B* dosage (*x*-axis). Data are from analysis of 6,748 SLE cases and 11,516 controls of European ancestry.



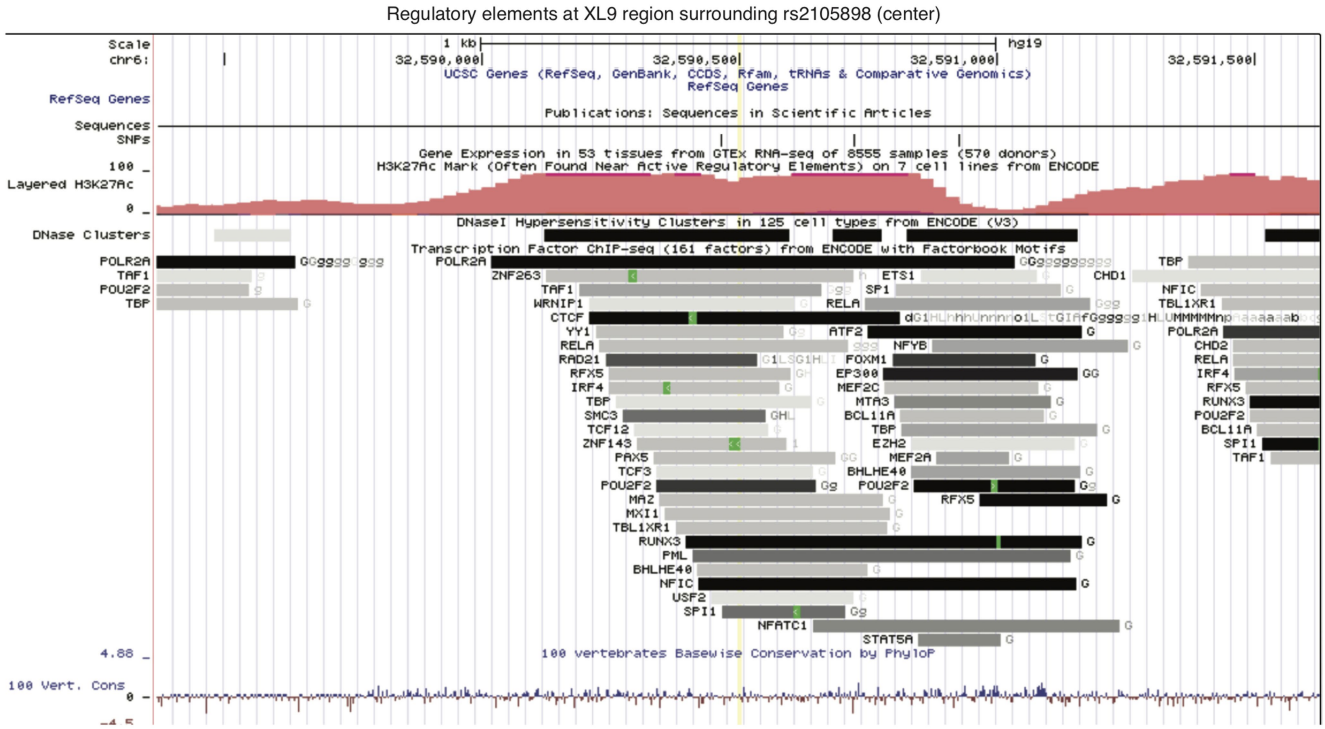
Extended Data Fig. 3 | Conditional association analyses for genetic markers across the extended MHC genomic region within the European-ancestry SLE and Sjögren's syndrome cohorts. a, Association of SLE with genetic markers (SNPs and imputed HLA alleles) across the extended MHC locus within the European-ancestry SLE cohort (6,748 cases and 11,516 controls). Orange diamond: an initial estimate of C4-related genetic risk, calculated as a weighted sum of the number of *C4A* and *C4B* gene copies: (2.3)*C4A*+*C4B*, with the weights derived from the relative coefficients estimated from logistic regression of SLE risk versus *C4A* and *C4B* gene dosages. This risk score is imputed with an accuracy (r^2) of 0.77. Points representing all other genetic variants in the MHC locus are shaded orange according to their level of LD-based correlation to this C4-derived risk score. **b**, As in **a**, but for a European-ancestry Sjögren's syndrome (SjS) cohort (673 cases and 1,153 controls). The orange diamond here also represents (2.3)*C4A* + *C4B*, with this weighting derived from the relative coefficients estimated from logistic regression of Sjögren's syndrome risk versus *C4A* and *C4B* gene dosages. **c**, Association of SLE with genetic markers (SNPs and imputed HLA alleles) across the extended MHC locus within the European-ancestry SLE cohort controlling for C4 composite risk (weighted sum of risk associated with various combinations of *C4A* and *C4B*). Variants are shaded in purple by their LD with rs2105898, an independent association identified from trans-ancestral analyses. **d**, As in **c**, but in association with a European-ancestry Sjögren's syndrome cohort. Here a simpler linear model of risk contributed by *C4A* and *C4B* was used instead of a weighted sum across all possible combinations.



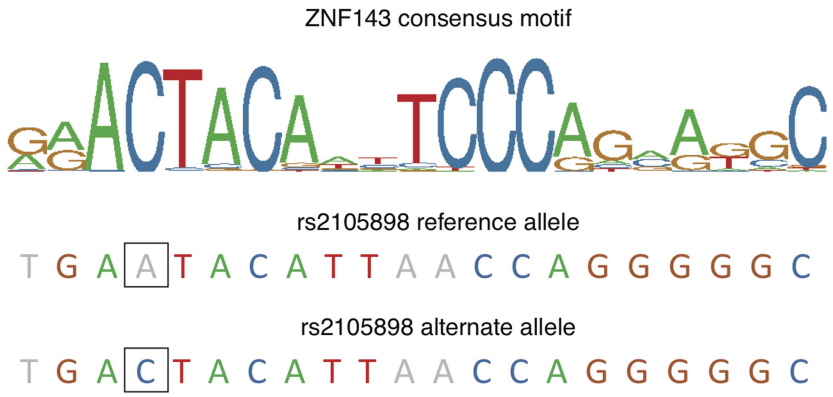
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Using C4 gene variation to understand the appearance of trans-ancestral disparity in MHC association signals, and to fine-map an additional genetic effect. Association signals (for SLE and Sjögren's syndrome) for variants in a multi-megabase region of human chromosome 6 containing the MHC region including the HLA and C4 genes. **a**, Relationship between SLE association ($-\log_{10}(p)$, *y*-axis) and LD to the weighted C4 risk score (*x* axis) for genetic markers and imputed HLA alleles across the extended MHC locus. In this European-ancestry cohort, it is unclear (from this analysis alone) whether the association with the markers in the predominant ray of points (at an angle of ~45° from the *x* axis) is driven by variation at C4 or by the long haplotype containing *DRB1**03:01 (green), *DQA1**05:01 (blue), *B**08:01 (red) and many other SNPs (black). In addition, at least one independent association signal (a ray of points at a higher angle in the plot, with strong association signals and only weak linkage disequilibirum-based correlation to C4 and *DRB1**0301) with some LD to *DRB1**15:01 (maroon) is also present. **b**, Analysis as in **a**, but for associations to Sjögren's syndrome in a cohort of European ancestry. As in SLE, it is initially unclear whether the genetic association signal is driven by variation at C4 or by linked HLA alleles, *DRB1**03:01 (green), *DQA1**05:01 (blue), and *B**08:01 (red). There is also an independent association signal with LD to *DRB1**15:01 (maroon). **c**, Analysis as in **a**, but of an African American SLE case–control cohort (in which LD in the MHC region is more limited). Many MHC-region SNPs associate with SLE in proportion to their LD with the weighted C4 risk score inferred from the earlier analysis of the European-ancestry cohort; this C4-derived risk score itself associates with SLE at $P = 4.3 \times 10^{-19}$ in a logistic regression on 1,494 SLE cases and 5,908 controls. No similarly strong association is observed for *DRB1**03:01, *DQA1**05:01 or *B**08:01, HLA alleles which are in strong LD with C4 risk on European-ancestry (but not African American) haplotypes. An independent association signal is also present in this cohort, more clearly in LD with the *DRB1**15:03 allele (maroon). **d**, LD in the European-ancestry SLE cohort between the composite C4 risk term (weighted sum of risk associated with various combinations of *C4A* and *C4B* from Fig. 2a) and variants in the MHC region as r^2 (*y*-axis). **e**, As in **d**, but for the African American SLE cohort. **f**, LD (to C4 composite risk) for the same variants in European-ancestry individuals (*x* axis) and African Americans (*y* axis). Note the abundance of variants that have greater LD with C4 risk among European-ancestry individuals than among African Americans. Also, several groups of variants have equivalent LD (to C4 risk) in European ancestry individuals but exhibit a range of LD to C4 risk among African Americans. **g**, Associations with SLE ($-\log_{10} P$ values) for the same variants in European ancestry (*x* axis) and African American (*y* axis) case–control cohorts. Orange shading represents the extent of LD with C4 risk in European ancestry individuals. Variants with strong European-specific association to SLE are generally in strong LD with C4 risk among European-ancestry individuals. **h**, Comparison of the inferred effect size from association of genetic markers with SLE (unconditioned log odds ratios) among European-ancestry (*x* axis) and African American (*y* axis) research participants. As also seen in **g**, variants with discordant associations to SLE (across populations) tend also to be in strong LD to C4 risk among European-ancestry individuals. **i**, As in **g**, but now controlling for the effect of C4 variation in analysis of the European-ancestry cohort (*x* axis). Note that controlling for C4 risk in European-ancestry individuals alone greatly aligns (relative to **g**) the patterns of association between European ancestry and African American cohorts. **j**, As in **i**, but now also controlling for the effect of C4 in associations of the African American cohort. Note that due to the lack of strong LD relationships between C4 and variants in the MHC region in African Americans (**e**), this further adjustment does not change results strongly (relative to **i**). The independent signal, rs2105898, and HLA alleles, *DRB1**15:01 and *DRB1**15:03, are also highlighted. LD with rs2105898 in European-ancestry individuals is indicated by purple shading. **k**, Comparison of the inferred effect sizes from association of genetic markers with SLE (log odds ratios) controlling for C4-derived risk among European-ancestry (*x* axis) and African American (*y* axis) research participants. Two SNPs (rs2105898 and rs9271513) that form a short haplotype common to both ancestry groups are among the strongest associations in both cohorts. (Their association to SLE in the European-ancestry cohort was initially much less remarkable than that of other SNPs that are in strong LD with C4.) LD with rs2105898 in European-ancestry individuals is indicated by purple shading. **l**, As in **i**, but with variants shaded by whether they exhibit greater LD to rs2105898 in Europeans (blue) or African Americans (red).

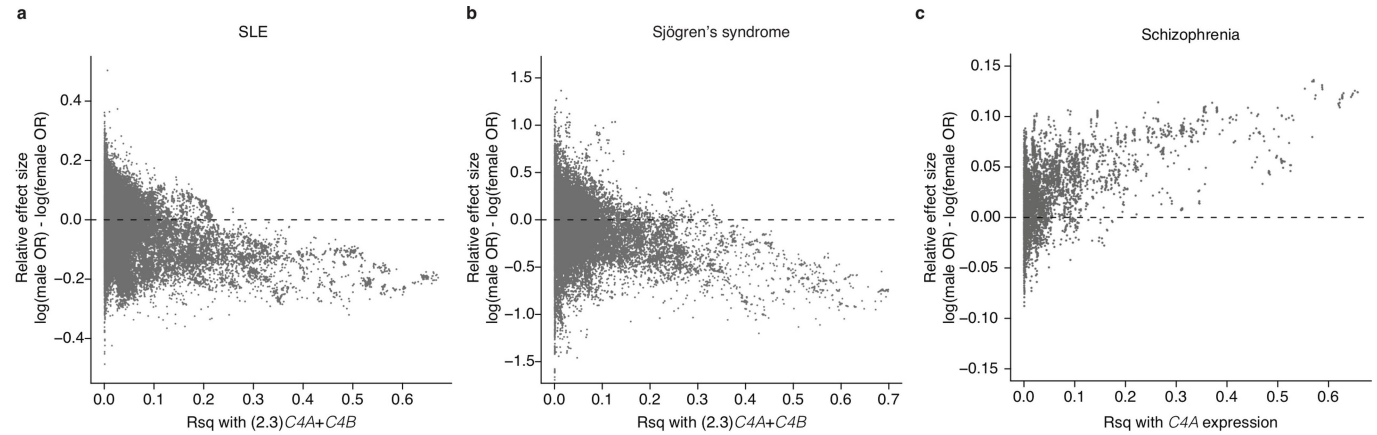
a



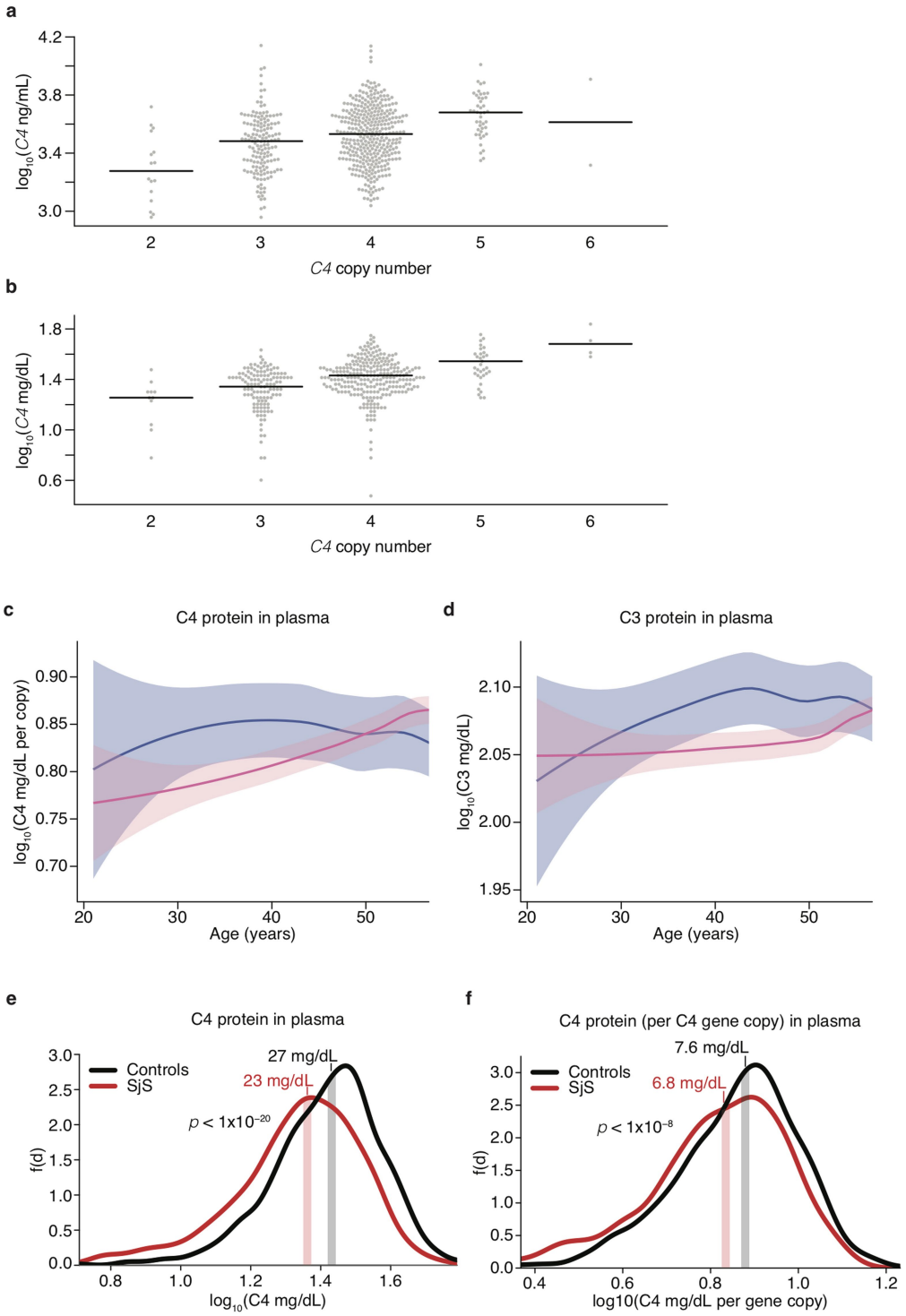
b



Extended Data Fig. 5 | Relationship of rs2105898 alleles to a known ZNF143 binding motif in the XL9 region of the MHC class II locus. a, Location of rs2105898 (yellow line at centre) within the XL9 region, with relevant tracks showing overlapping histone marks and transcription factor binding peaks (from ENCODE[52]), visualized with the UCSC genome browser[53]. **b**, ZNF143 consensus binding motif as a sequence logo, with the letters coloured if the base is present in more than 5% of observed instances. The alleles of rs2105898 are indicated by outlined box surrounding the base.



Extended Data Fig. 6 | Relationships between sex bias of disease associations and LD to C4 risk for variants in the MHC region. **a**, Relationship between male bias in SLE risk (difference between male and female log–odds ratios) and LD with C4 risk for common (minor allele frequency (MAF) > 0.1) genetic markers across the extended MHC region (6,748 cases and 11,516 controls). For each SNP, the allele for which sex risk bias is plotted is the allele that is positively correlated (via LD) with C4-derived risk score. **b**, Relationship between male bias in Sjögren's syndrome risk (log-odds ratios) and LD with C4 risk for common (MAF > 0.1) genetic markers across the extended MHC region (673 cases and 1,153 controls). For each SNP, the allele for which sex risk bias is plotted is the allele that is positively correlated (via LD) with C4-derived risk score. **c**, Relationship of male bias in schizophrenia risk (log odds ratios) and LD to *C4A* expression for common (MAF >0.1) genetic markers across the extended MHC region (28,799 cases and 35,986 controls). For each SNP, the allele for which sex risk bias is plotted is the allele that is positively correlated (via LD) with imputed *C4A* expression, as previously described[7].



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Correlation of C4 protein measurements in CSF and blood plasma with imputed C4 gene copy number and relationship of plasma complement to sex and Sjögren's syndrome status. **a**, Measurements of C4 protein in CSF obtained by ELISA ($n = 507$ total) are presented as $\log_{10}$[concentration (ng ml^{-1})] (y axis) for each observed or imputed copy number of total C4 (x axis, here showing most likely copy number from imputation). Because C4 gene copy number affects C4 protein levels so strongly, we normalized C4 protein measurements to each donor's C4 gene copy number in subsequent analyses (Fig. 3c). Bars indicate median values for each C4 copy number. **b**, Measurements of C4 protein in blood plasma obtained by immunoturbidimetric assays are presented as $\log_{10}$[concentration (mg dl^{-1})] (y axis) for each imputed most-likely copy number of C4 genes (x axis). Because C4 gene copy number affects C4 protein levels so strongly, we normalized C4 protein measurements by C4 gene copy number in subsequent analyses as in **c**. Due to the number of observations ($n = 1{,}844$ total), the plot is downsampled to 500 points; the median bars shown are for all individuals (before downsampling). **c**, Levels of C4 protein in blood plasma from 182 adult men and 1,662 adult women as a function of age. Concentrations are normalized to the number of C4 gene copies in an individual's genome (a strong independent source of variance) and shown on a $\log_{10}$ scale as a LOESS curve. Shaded regions represent 95% confidence intervals derived during LOESS. **d**, Levels of C3 protein in blood plasma as a function of age from the same individuals in panel **c**. Concentrations are shown on a $\log_{10}$ scale as a LOESS curve. Shaded regions represent 95% confidence intervals derived during LOESS. **e**, C4 protein in blood plasma was measured in 670 individuals with Sjögren's syndrome (red) and 1,151 individuals without Sjögren's syndrome (black) and is shown on a $\log_{10}$ scale (x axis). Vertical stripes represent median levels for cases and controls separately. Comparison of the two sets was done with a non-parametric two-sided Mann–Whitney rank-sum test ($P = 4.8 \times 10^{-21}$). **f**, As in **e**, but concentrations are normalized to the number of C4 gene copies in an individual's genome; this per-copy amount is shown on a $\log_{10}$ scale (x axis). Comparison of the two sets was done with a non-parametric two-sided Mann–Whitney rank-sum test ($P = 7.6 \times 10^{-9}$).

Extended Data Table 1 | Imputation accuracy for C4 copy numbers in European ancestry and African American haplotypes

	Imputation accuracy (r^2)	
Gene copy number	**European ancestry**	**African Americans**
C4	0.80	0.58
C4A	0.78	0.65
C4B	0.74	0.61
C4-HERV	0.91	0.76
2.3(*C4A*)+*C4B*	0.77	0.64

Imputation accuracy was evaluated by correlation of imputation results to C4 gene copy numbers directly inferred from WGS data. Aggregated copy numbers imputed from each round of leaving ten individuals out were correlated with the directly-typed measurements and are reported as r^2 for each feature of C4 structural variation for European ancestry (693 individuals) and African American (250 individuals) members of the reference panel separately.

Extended Data Table 2 | Frequency of common C4 alleles and their LD-based correlation with HLA alleles in European ancestry and African American cohorts

European ancestry

A			*B*			*C*			*C4* allele	Allele Frequency	*DRB1*			*DQA1*			*DQB1*		
allele	%	r^2	allele	%	r^2	allele	%	r^2			allele	%	r^2	allele	%	r^2	allele	%	r^2
01:01	69	0.27	08:01	93	**0.75**	07:01	93	**0.57**	B(S)	13.7%	03:01	94	**0.71**	05:01	94	**0.7**	02:01	94	**0.7**
									A(L)	4.8%									
						06:02	69	0.31	A(L)-B(S)-1	6.1%	07:01	74	0.25	02:01	74	0.25			
			44:03	54	0.28	16:01	53	0.39	A(L)-B(S)-2	4.5%	07:01	57	0.1	02:01	57	0.1	02:02	55	0.14
									A(L)-B(S)-3	3.8%									
									A(L)-B(S)-4	4.5%									
			07:02	64	**0.42**	07:02	63	0.35	A(L)-B(L)-1	15.5%	15:01	73	**0.49**	01:02	74	0.32	06:02	70	**0.47**
									A(L)-B(L)-2	23.1%									
			35:01	55	0.2	04:01	57	0.09	A(L)-A(L)-1	3.2%	01:01	65	0.14	01:01	65	0.11	05:01	64	0.1
									A(L)-A(L)-2	2.1%	13:01	67	0.16	01:03	65	0.13	06:03	67	0.15
02:01	65	0.03	44:02	74	0.24	05:01	72	0.23	A(L)-A(L)-3	4.5%	04:01	80	0.29	03:03	79	0.37	03:01	82	0.15

African American

A			*B*			*C*			*C4* allele	Allele Frequency	*DRB1*			*DQA1*			*DQB1*		
allele	%	r^2	allele	%	r^2	allele	%	r^2			allele	%	r^2	allele	%	r^2	allele	%	r^2
									B(S)	5.0%				01:02	51	0.01			
									A(L)	7.5%									
									A(L)-B(S)-1	14.1%									
									A(L)-B(S)-2	18.1%									
									A(L)-B(S)-3	17.7%									
									A(L)-B(S)-4	6.5%									
									A(L)-B(L)-1	4.4%	15:01	67	0.2	01:02	72	0.04	06:02	59	0.06
									A(L)-B(L)-2	4.5%									
									A(L)-A(L)-1	0.7%	01:01	57	0.07	01:01	53	0.01			
									A(L)-A(L)-2	0.8%									
02:01	72	0.03	44:02	86	0.31	05:01	78	0.17	A(L)-A(L)-3	0.8%	04:01	93	0.27	03:03	86	0.14	03:01	87	0.03

For each common C4 allele and HLA gene, the allele with strongest LD (r^2) is listed if present on more than half of the haplotypes with that C4 allele (for 36,528 European ancestry and 14,804 African American haplotypes separately, with exact fraction as a percentage). r^2 values greater than 0.4 are highlighted to point out particularly strong C4-HLA allele correlations, such as for several HLA alleles with the C4-B(S) allele in European ancestry individuals. Some common C4 alleles are further subdivided into distinct haplotypes used in imputation (and in Fig. 1b), as defined by shared alleles from variants flanking C4. Note that some alleles such as C4-A(L)-A(L)-3 are present at a low frequency in African Americans that might reflect their presence on admixed European-origin haplotypes spanning this region, whereas others such as C4-B(S) are likely to also exist on African haplotypes – these differences between C4 alleles are also reflected in the similarity of LD with HLA alleles to the corresponding row of the European ancestry section.

Extended Data Table 3 | Results of association analyses of SLE risk against C4 variation, HLA alleles, and/or rs2105898 in European ancestry and African American cohorts

European ancestry

	C4			*C4A*			*C4B*			*DRB1**03:01			*B**08:01			rs2105898				
Model	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	AIC	LRT -log10(p)
C4	-0.55	0.027	92.7																22855.26	260.2
C4A				-0.53	0.024	105.3													22790.05	274.3
C4A+C4B				-0.62	0.028	112	-0.27	0.037	12.3										22739.8	284.4
*DRB1**03:01										0.7	0.03	117.1							22748.33	283.3
*B**08:01													0.69	0.031	108.4				22790.65	274.2
rs2105898																-0.32	0.027	30.7	23153.86	195.5
*C4A + C4B + DRB1**03:01				-0.35	0.041	17.2	-0.11	0.041	2.3	0.4	0.046	17.5							22666.1	299.6
*C4A + C4B + B**08:01				-0.41	0.039	24.6	-0.17	0.039	4.7				0.35	0.044	14.4				22680.53	296.4
C4A + C4B + rs2105898				-0.67	0.028	122.8	-0.32	0.038	16.4							-0.38	0.028	41.1	22558.42	322.8

African American

	C4			*C4A*			*C4B*			*DRB1**03:01			*B**08:01			rs2105898				
Model	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	beta	se	-log10(p)	AIC	LRT -log10(p)
C4	-0.51	0.059	17.3																7358.65	19.7
C4A				-0.43	0.062	11.2													7385.17	14
C4A+C4B				-0.62	0.068	18.7	-0.41	0.068	8.6										7351.45	20.9
*DRB1**03:01										0.41	0.091	5.2							7413.36	8
*B**08:01													0.78	0.11	11.6				7387.33	13.6
rs2105898																-0.46	0.047	21.9	7339.35	23.9
*C4A + C4B + DRB1**03:01				-0.59	0.073	15	-0.38	0.071	7.1	0.1	0.099	0.5							7352.34	20.4
*C4A + C4B + B**08:01				-0.51	0.073	11.7	-0.37	0.069	7.2				0.49	0.12	4.4				7337.24	23.6
C4A + C4B + rs2105898				-0.52	0.07	13.2	-0.43	0.069	9.4							-0.42	0.048	17.8	7277.78	36.2

Coefficients (beta, standard error) and *P* values (as $-\log_{10}(P)$) for individual terms composing several relevant logistic regression models for predicting SLE risk in a European ancestry cohort of 6,748 SLE cases and 11,516 controls and an African American cohort of 1,494 SLE cases and 5,908 controls. Each analysis also included ancestry-specific covariates. For each model, the Akaike information criterion (AIC) and overall *P* value (as determined by Chi-squared likelihood-ratio test) are given on the right to indicate the relative strengths of similar models for each ancestry cohort.

nature research

Corresponding author(s): Steven A. McCarroll

Last updated by author(s): Feb 19, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	BioTek ELx800 Gen 5 software was used to collect ELISA absorbance readout on microplates for CSF samples.
Data analysis	Genome STRiP 2.0 was used for C4 copy number calling on whole genome-sequenced samples. BEAGLE v4.1 (21Jan17.6cc) was used for imputation of C4 variation. HIBAG v1.4 was used for HLA allele imputation. R v3.6 was used for downstream analyses and functions were derived largely from default packages (e.x. stats) with the exception of third-party HIBAG and ggplot2 packages.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Individual genotype data for Sjögren's syndrome cases and controls and individual plasma concentrations for C4 and C3 are available in dbGaP under accession number phs000672.v1.p1. Individual genotype data for schizophrenia cases and controls are available by application to the Psychiatric Genomics Consortium (PGC). Questions regarding individual genotype data for SLE cases and controls of European and/or African American ancestry can be directed to Timothy J. Vyse (timothy.vyse@kcl.ac.uk). Data resources (reference haplotypes), software scripts and instructions for imputing C4 alleles into SNP data sets are available at http://mccarrolllab.org/resources/resources-for-c4/. C4 genotype and protein concentration data for CSF samples are available upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For genetic analyses, we were maximally inclusive of human-genetic datasets available at the time of analyses, and collaborated internationally to achieve the largest sample sizes we could: e.g. 6,748 SLE cases and 11,516 controls of European ancestry; 1,494 SLE cases and 5,908 controls for African Americans. A strong pre-analysis indicator that these sample sizes would be sufficient, came from the fact that earlier work on these same data sets had already established extremely strong associations to genetic markers at the MHC locus ($p < 1e-100$ among Europeans; $p < 1e-25$ among African Americans). For analyses of the relationship of CSF complement protein levels to sex and age, we sampled from a larger panel of CSF samples so as to include sufficient numbers of samples within the age ranges (20-50) that correspond to sex-biased disease incidence. We used sample sizes that were comparable to or larger than those in previous CSF studies. Evidence that these sample sizes were sufficient came from the strong statistical significance of the results.
Data exclusions	For human-genetic analyses, pre-established QC metrics standard in the field were used to exclude some samples or genotypes for analysis, as described in Methods; these were pre-established criteria similar to those used in most human genetics studies. These included: (i) exclusion of SNPs based on genotyping rate and Hardy-Weinberg equilibrium; (ii) relatedness (genotyped individuals were excluded if we found them to be related to one another, based on predetermined cutoffs for relatedness, such as excluding duplicate samples and close relatives); (iii) any disagreements of annotated characteristics (such as sex or ancestry) with the inference of these same characteristics from genotype data. It was also pre-determined (before ELISA assays) that CSF samples were to be excluded if they had any visually apparent blood contamination.
Replication	Genetic findings were first critically evaluated by analyses finding that results were consistent across two distinct levels of analysis: (i) the copy number of C4A and C4B genes (Fig. 1a); and (ii) the haplotypes formed by C4 structural alleles and flanking SNPs (Fig. 1b). We then replicated the results for SLE by an independent analysis in another cohort. We found that the findings on C4-associated risk levels were consistent (Fig. 2a) across populations (European-ancestry and African American research cohorts) with different ancestries and different patterns of linkage disequilibrium. We further replicated these results by finding the results to be consistent with those in an independent cohort of patients with a closely related illness (Sjogren's, Fig. 1b). Finally, one of the most surprising results (the finding that C4 alleles associated with larger effects in men) was replicated in a distinct illness, schizophrenia (Fig. 3ab). For analyses of complement protein concentrations in men and women, we analyzed two panels of CSF samples which had been collected by different investigators at different hospitals. We found that the finding of sex bias (higher levels in males than females) was consistent across these cohorts and significant in each cohort independently. We also replicated the CSF results in plasma by re-analyzing data from an earlier study.
Randomization	Individuals genotyped for disease associations had been previously organized into cohorts (with matched controls) by disease status and ancestry. SNP genotyping was done in batches, as described in the original publications in which the SNP genotype data were generated. To address the possibility that population stratification or batches could contribute to any results, we utilized a practice (standard in well-powered human-genetic studies that have access to genome-wide SNP data) of addressing such potential influences by calculating the principal components (PCs) of the genotype matrix for each cohort, then using the PC scores as covariates in logistic-regression association analysis. For schizophrenia analyses, for which multiple cohorts of European ancestry had been collected, the sample's collection site was encoded as an additional indicator covariate in logistic regression, to control for variability in diagnostic thresholds.
Blinding	Blinding was accomplished by the use of an ID number for each sample, which was only re-associated with metadata (e.g. donor sex) in the final statistical analysis. Thus, for example, laboratory analyses of CSF proteins occurred in a manner blinded to donor characteristics.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Polyclonal Antiserum to Human C4 Protein; supplier: Quidel; catalog number: A305; lot number: 903556-1; dilution: 1:1000
Polyclonal Rabbit Anti-Human C4c Complement/FITC; supplier: Dako; catalog number: F016902-2; lot number: 89152; dilution: 1:3000
Goat Anti-Rabbit IgG H&L (Alkaline Phosphatase); supplier: abcam; catalog number: ab97048; lot number: GR166802-2; dilution: 1:5000
Human Complement C3 ELISA Kit; supplier: abcam; catalog number: ab108823

Validation

All antibodies are validated as described in their respective technical data sheets or similar; these statements along with citations can be found on supplier webpages for the product. We also quote highlights from these documents here. For polyclonal antiserum to human C4 protein (Quidel, A305), "Highly purified human C4 was isolated from normal serum and used to immunize goats. The anti-human C4 polyclonal antisera was tested against normal human plasma by double immunodiffusion, one-dimensional immunoelectrophoresis,quantitative radial immunodifussion, and quantitative rocket immunoelectrophoresis. The antiserum was determined to be monospecific for C4 at varying concentrations. Applications of the C4 polyclonal antisera have been evaluated by various research facilities, and include, Western Blot, IHC, Immunofluorescence, and ELISA." For polyclonal rabbit anti-human C4c Complement/FITC (Dako, F016902-2), "The antibody reacts with C4, C4b and C4c, but does not react with the C4d fragment. Traces of contaminating anti-bodies have been removed by solid-phase absorption with human plasma proteins. The specificity has been ascertained as follows:Crossed immunoelectrophoresis:Only reactivity with C4 complement and its C4c-containing fragments is observed when using unconjugated antibody corresponding to 40 uL F-0169 per square cm gel area against 2 uL human plasma. Staining: Coomassie Brilliant Blue. In rocket immunoelectrophoresis the antibody cross-reacts with C4c complement from all of 11 animal species tested so far: Cat, cow, dog, goat, guinea pig, horse, mink, mouse, rat, sheep and swine."

Human research participants

Policy information about studies involving human research participants

Population characteristics

Our analyses included patients of both sexes and multiple ancestries (European and African American). These cohorts have been described in previously published studies (cited in the current work); we summarize their most analysis-relevant characteristics here. We addressed the effects of cryptic (unseen) ancestry by calculating principal components of the genotype matrix and using these as additional covariates in association analysis; this is standard practice in well-powered human genetics studies that have access to genome-wide data. Additional key covariates included sex (men comprised 27% of the European-ancestry SLE cohort, 29% of African American SLE cohort, 10% of the SjS cohort, and 61% of the schizophrenia cohort), collection site/cohort (used in schizophrenia analyses, to account for variation in diagnostic thresholds and ascertainment strategies at sites contributing data to the Psychiatric Genomics Consortium analyses; this was encoded for analysis as a set of indicator variables for membership in each of 40 cohorts), and smoking status (used in Sjögren's syndrome analysis; 12% of the cohort were current smokers). For each disease, genotyped case-control cohorts were as described in prior publications (cited in the current work), in which detailed definitions of phenotypes and associated covariates can be found. Relevant metadata for plasma samples were age (22-89 years old), sex (10% men), and disease status (670 patients were clinically diagnosed with Sjögren's syndrome by meeting the American College of Rheumatology classification criteria for Sjögren's syndrome) and were as described in the dbGaP study with accession number phs000672.v1.p1. CSF sample metadata of age (18-64 years old) and sex (67% men) were recorded upon collection.

Recruitment

For previously-collected samples – including genomic DNA for genotyping (from >40 sites), plasma complement measurements, and one CSF sample panel – recruitment was as described in the previously published studies (cited in the current work). For one set of CSF samples that has not been described in previous papers, CSF was drawn in a hospital context to evaluate for the possibility of CNS infection (cases of confirmed infection were excluded from the collection).

Ethics oversight

Statistical analyses at Harvard Medical School received an NHSR determination from the Harvard Medical School IRB. For previously-collected samples – including genomic DNA for genotyping (from >40 sites), plasma complement measurements, and one CSF sample panel – local IRBs at each institution had approved the collections and patient-consent materials, as described in the earlier papers on these cohorts (cited in the current work). For one set of CSF samples that has not been described in previous papers, the IRB at Brigham and Womens Hospital approved this under protocol #1999P010911.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Cellular locomotion using environmental topography

https://doi.org/10.1038/s41586-020-2283-z

Received: 13 January 2018

Accepted: 9 March 2020

Published online: 13 May 2020

Check for updates

Anne Reversat[1,7]✉, Florian Gaertner[1], Jack Merrin[1], Julian Stopp[1], Saren Tasciyan[1], Juan Aguilera[1], Ingrid de Vries[1], Robert Hauschild[1], Miroslav Hons[1,2,8], Matthieu Piel[3,4], Andrew Callan-Jones[5], Raphael Voituriez[6] & Michael Sixt[1]✉

Eukaryotic cells migrate by coupling the intracellular force of the actin cytoskeleton to the environment. While force coupling is usually mediated by transmembrane adhesion receptors, especially those of the integrin family, amoeboid cells such as leukocytes can migrate extremely fast despite very low adhesive forces[1]. Here we show that leukocytes cannot only migrate under low adhesion but can also transmit forces in the complete absence of transmembrane force coupling. When confined within three-dimensional environments, they use the topographical features of the substrate to propel themselves. Here the retrograde flow of the actin cytoskeleton follows the texture of the substrate, creating retrograde shear forces that are sufficient to drive the cell body forwards. Notably, adhesion-dependent and adhesion-independent migration are not mutually exclusive, but rather are variants of the same principle of coupling retrograde actin flow to the environment and thus can potentially operate interchangeably and simultaneously. As adhesion-free migration is independent of the chemical composition of the environment, it renders cells completely autonomous in their locomotive behaviour.

Mechanistic understanding of cell motility started with Abercrombie's demonstration that particles placed on the dorsal surface of a cell undergo retrograde transport[2]. The authors concluded that cellular material is added to the front of the cell, feeding a rearwards flow, which couples to the substrate and thereby generates friction to drag the cell body forwards. Later studies showed that the added material is the growing polymer of F-actin and that F-actin retrograde flow couples via transmembrane adhesion receptors, mainly of the integrin family. Although this adhesion-dependent principle of motility turned out to be universal on two-dimensional (2D) surfaces, some cells can migrate in the absence of integrin-mediated adhesion when confined within a three-dimensional (3D) context[3,4]. This applies especially to the amoeboid (shape-changing) class of cells such as leukocytes[5]. Force transmission under these conditions is not understood, and alternative membrane-spanning molecules that mediate specific or even unspecific friction have been suggested[6,7]. Notably, in his seminal study, Abercrombie contemplated a second, qualitatively different, scenario: backward-moving waves of deformation might passively entangle with the applied particles (or a substrate) and propel the cell in a manner analogous to paddling and swimming[2].

To explore adhesion-independent locomotion, we initially investigated T lymphocytes. When these cells migrate within lymphatic organs, integrin depletion causes them to slow down by only 15%, meaning that they do employ integrins but retain a substantial capacity to transmit forces via other means[8,9]. We used a mouse T cell line, which migrates vigorously in 3D environments, such as in vitro-assembled collagen gels (Supplementary Video 1). To eliminate integrin-based force transmission, we deleted *Tln1*, which encodes talin 1, a cytosolic adaptor protein essential for integrin functionality[10] (Extended Data Fig. 1a–c). Talin-deficient (referred to here as 'talin KO') cells were completely unable to adhere to and migrate on 2D surfaces (Fig. 1a, b, Extended Data Fig. 1d, e). However, once incorporated into 3D collagen gels, their migratory characteristics were indistinguishable from wild-type cells (Fig. 1c, Extended Data Fig. 1f, Supplementary Video 1). In the following, we engineered artificial environments to dissect the geometrical transition between 2D surfaces and fibrillar environments, while keeping the chemical composition of the substrate unchanged. We first designed a microfluidic setup to confine leukocytes between two parallel surfaces separated by an adjustable space[11] ('2.5D'; Fig. 1d). While this setup supported efficient migration of wild-type cells, talin-KO cells were unable to translocate, although they showed morphological polarization (Fig. 1d, Extended Data Fig. 1g, Supplementary Video 2). When T cells expressing the actin reporter Lifeact-GFP were imaged with total internal reflection fluorescence (TIRF) microscopy, actin within the rapidly moving cell body remained static in relation to the adhesive substrate and flowed backwards in the cell frame of reference (Fig. 1e–g, Supplementary Video 3). Talin-KO cells on adhesive substrates or wild-type cells on passivated surfaces were immobile but showed similar actin dynamics in the cell frame of reference[8], meaning that actin slid backwards in relation to the substrate (Fig. 1e–g, Extended Data Figs. 1g–l, 2a–c, Supplementary Video 3). These findings demonstrate that, when placed on 2D substrates or when confined

[1]Institute of Science and Technology Austria (IST Austria), Klosterneuburg, Austria. [2]Institute of Scientific Instruments of the Czech Academy of Sciences, Brno, Czech Republic. [3]Institut Curie, PSL Research University, CNRS, UMR 144, Paris, France. [4]Institut Pierre-Gilles de Gennes, PSL Research University, Paris, France. [5]Laboratoire Matière et Systèmes Complexes, UMR 7057 CNRS, Université Paris Diderot, Paris, France. [6]Laboratoire de Physique Theorique de la Matière Condensée et Laboratoire Jean Perrin, CNRS/Université Pierre-et-Marie Curie, Paris, France. [7]Present address: Institute of Translational Medicine, Department of Cellular and Molecular Physiology, University of Liverpool, Liverpool, UK. [8]Present address: BIOCEV, First Faculty of Medicine, Charles University, Vestec, Czech Republic. ✉e-mail: anne.reversat@gmail.com; sixt@ist.ac.at

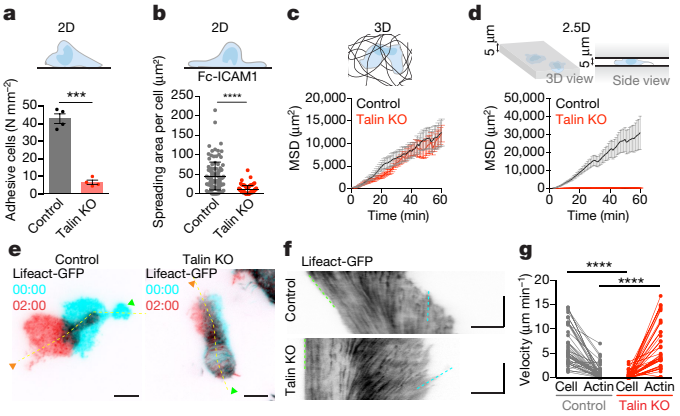


Fig. 1 | Adhesive and migratory properties of talin-KO T cells. a, The number of cells adhering to the 2D surface. Data are mean ± s.e.m. of four independent experiments; ***$P = 0.00076$, paired *t*-test. **b**, Surface area (μm²) of cells spreading on Fc-ICAM1-coated glass. Control ($n = 139$) and talin-KO ($n = 81$) cells of three independent experiments are shown. Data are mean ± s.d.; ****$P < 0.0001$, Mann–Whitney *U*-test. **c**, Mean square displacement (MSD) of control ($n = 51$, in black) and talin-KO ($n = 52$, in red) cells migrating in 3D collagen gel from three independent experiments. Data are mean ± s.e.m.; $P = 0.2243$, Mann–Whitney *U*-test. **d**, MSD of control ($n = 290$, in black) and talin-KO ($n = 284$, in red) cells migrating in a 5-μm confiner from four independent experiments. Data are mean ± s.e.m.; $P < 0.0001$, Mann–Whitney *U*-test. **e**, Snapshots of TIRF microscopy of a control cell (left) and a talin-KO cell (right) expressing Lifeact-GFP under 5-μm confinement. Representative of three independent experiments. The green arrowheads mark the uropod and the orange arrowheads indicate the cell front. TIRF images in cyan are at $t = 0$ min and in red at $t = 2$ min. The yellow dashed line is used for the kymograph in **f**. Scale bars, 5 μm. Time is shown in min:s. **f**, Kymographs of a control cell and talin-KO cell under 5-μm confinement as shown in **e**. Cell velocity is indicated by the green dashed line, and actin retrograde flow is shown by the cyan dashed line. Horizontal scale bar, 5 μm; vertical scale bar, 1 min. **g**, Retrograde flow velocities for control ($n = 42$) and talin-KO ($n = 36$) cells from three independent experiments. ****$P < 0.0001$, Mann–Whitney *U*-test.

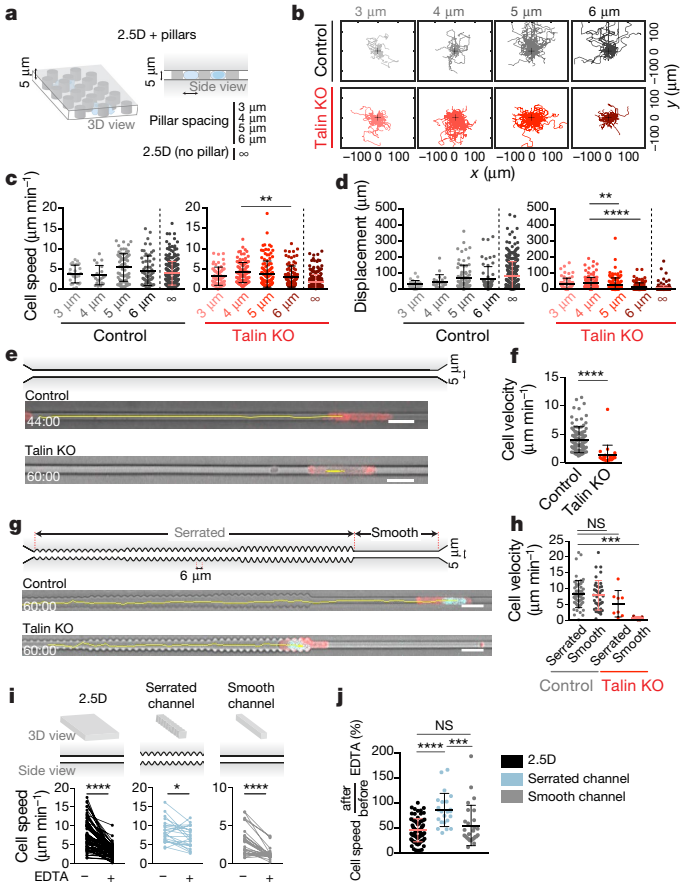


Fig. 2 | Migration of T cells in textured confinement. a, Experimental setup for **b**–**d**. **b**, Cell trajectories analysed in **c** and **d** during 1 h. The black crosses indicate the start of the individual tracks. **c**, **d**, Cell speed (**c**), mean ± s.d.; **$P = 0.0021$, otherwise not significant ($P > 0.9999$); and displacement (**d**), mean ± s.d.; **$P = 0.0089$ and ****$P < 0.0001$, otherwise not significant, Kruskal–Wallis test followed by post hoc Dunn's test. $n = 25$, 20, 75 and 61 for control cells in respective 3-μm, 4-μm, 5-μm and 6-μm zones. Data are from three independent experiments. $n = 64$, 115, 124 and 94 for talin-KO cells in respective 3-μm, 4-μm, 5-μm and 6-μm zones. Data are from six independent experiments. **e**, **f**, Migration in smooth channels. Data are from four independent experiments. In **e**, 5 × 5-μm smooth channels are shown. Top, scheme of the channel. Representative snapshots at $t = 44$ min of Lifeact-GFP-expressing control cells (middle) and at $t = 60$ min talin-KO cells (bottom) are displayed. Tracks are shown in yellow. Scale bars, 20 μm. In **f**, the cell velocity is shown. $n = 99$ for control cells, $n = 24$ for talin-KO cells. Data are mean ± s.d.; ****$P < 0.0001$, Mann–Whitney *U*-test. **g**, **h**, Migration in serrated channels. Data are from three (control) and four (KO) independent experiments. In **g**, the serrated topography with a 6-μm period in a 5 × 5-μm channel is shown. Top, scheme of the channel. A representative snapshot at $t = 60$ min of the Lifeact-GFP control cell (middle) and the talin-KO cell (bottom) is displayed. Tracks are shown in yellow. Scale bars, 20 μm. In **h**, the cell velocity in channels is shown. $n = 59$ (control serrated), 25 (control smooth), 8 (talin-KO serrated) and 5 (talin-KO smooth). Data are mean ± s.d.; ***$P = 0.0009$, Kruskal–Wallis test followed by post hoc Dunn's test. **i**, **j**, Cell speed before and after the addition of 10 mM EDTA. Data are from three independent experiments. 2.5D: $n = 73$ cells; serrated channel: $n = 23$; and smooth channel: $n = 27$. In **i**, the cell speed before (–) and after (+) the addition of EDTA in respective devices is shown. ****$P < 0.0001$ and *$P = 0.0167$, Wilcoxon matched-pairs signed-rank test. In **j**, the change in cell speed after versus before the addition of EDTA is shown. Data are mean ± s.d.; not significant (NS; $P > 0.9999$), ***$P = 0.0004$ and ****$P < 0.0001$, Kruskal–Wallis test followed by post hoc Dunn's tests.

between two surfaces, T cell locomotion depends on integrin-mediated force transmission[6,8,9,12] (see also Supplementary Discussion). This finding was in contrast to our collagen gel data and to the in vivo finding that T cells retain substantial migratory capacity following integrin depletion[8,9]. We reasoned that the major difference between confinement within a lymph node or within a fibrillar matrix and confinement between two planar surfaces is geometrical complexity. Hence, we introduced arrays of variably spaced pillars that intersected the two surfaces (Fig. 2a), thereby mimicking the geometry of a fibrillar gel while keeping the surface chemistry unchanged[13]. Between the pillars, the migratory capacity of talin-KO cells was restored (Fig. 2b–d, Extended Data Fig. 2d, e, Supplementary Video 4). While translocation velocities of wild-type cells were not affected by increased pillar spacings, the performance of talin-KO cells dropped in wider-spaced pillars, suggesting that tight contact with curved surfaces enables propulsion in the absence of integrins (Supplementary Video 4). We next turned to channels in which cells are fully enclosed by surfaces (Fig. 2e). Within smooth-walled 5 × 5-μm channels, control cells migrated persistently, while talin-KO cells were immobile (Fig. 2e, f, Supplementary Video 5). We next introduced surface texture as a serrated pattern with a 6-μm period and found that the migration of talin-KO cells was restored (Fig. 2g, h, Extended Data Fig. 2f, Supplementary Video 5). In channels that contained both smooth and serrated sectors, wild-type cells transited freely between the sectors. Following cation depletion with EDTA, which inactivates integrins (and several other cell-surface adhesion molecules), cells within smooth areas slowed down or stalled, while cells within serrated sectors kept migrating until they encountered a smooth area (Fig. 2i, j, Supplementary Video 6). The same principle was conserved when we tested dendritic cells and neutrophil granulocytes, two other types of amoeboid leukocytes. Notably, EDTA-treated

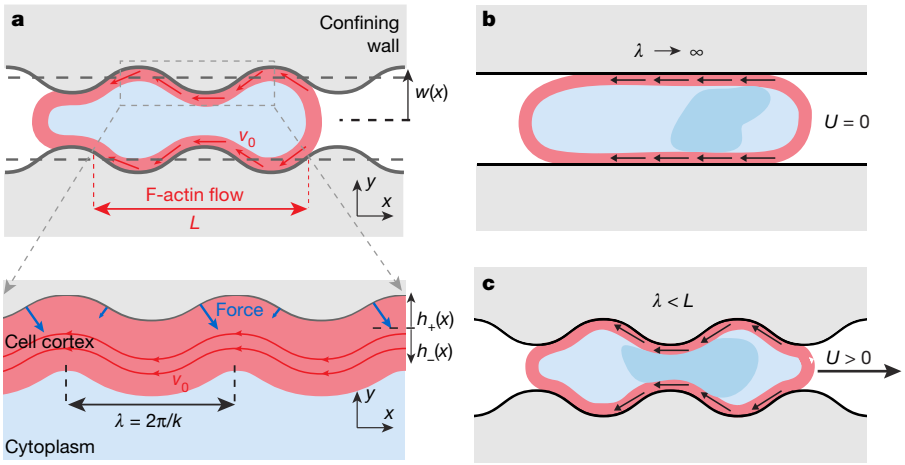


Fig. 3 | Physical model of adhesion-free cell migration in a complex environment. a, Scheme of the model: the channel has a varying wavy cross-section $w(x)$ and period $\lambda = 2\pi/k$. The cell contains cortical F-actin that presents a retrograde flow of length L, an average speed v_0 and is modelled as a viscous fluid layer of thickness $h(x)$. **b**, Scheme of a modelled cell in a smooth channel with an adhesion-free coupling of the actin cytoskeleton with the environment. The cell is polarized and has a high retrograde actin flow (black arrows) but cannot move forwards. **c**, Scheme of the modelled non-adhesive cell in a serrated channel. Cell propulsion is created by the shear stress in the actin cortex, which is caused by the bending of flow lines (black arrows) by the environmental topography. This induces a pressure gradient in the cortical F-actin mesh and thus non-vanishing normal forces onto the substrate. U, cell speed.

dendritic cells retained a weak locomotive capacity even in smooth channels (Extended Data Fig. 3a, b, Supplementary Video 7). This is in line with previous findings that, compared with T cells, dendritic cells have a higher capacity to transmit forces via unspecific transmembrane coupling[8,12,14]. These data demonstrate that force transmission of leukocytes follows a continuum of strategies. They can flexibly use integrin-mediated adhesion, unspecific transmembrane force coupling and, if both strategies fail due to the absence of suitable ligands, can efficiently utilize surface topography to transmit forces.

As a framework to understand topography-based locomotion, we propose a minimal model (Fig. 3a–c, Supplementary Text) in which actin is described as a viscous gel that travels from the leading to the trailing edge of the cell in the absence of any tangential friction force with the substrate. In serrated but not in smooth channels, this flow generates local shear stresses and thus a pressure gradient in the actin gel, which drives locomotion by imposing non-vanishing normal forces onto the substrate. Here the actin flow generated in the cell can couple to the environment when its topographical features are smaller than the flow scale of the actin cortex. To test this prediction, we varied the period of serrated topography from 6 to 12 to 24 µm. Serrated sectors were followed by smooth areas as an internal control (Fig. 4a). Wild-type T cells effectively traversed all (including smooth) channel designs, while talin-KO cells migrated in 6-µm and 12-µm period patterns but rarely in the 24-µm patterns (Fig. 4b–d, Extended Data Fig. 4a), in which the spacing exceeded the average length scale of cortical actin flow in a cell (Extended Data Fig. 4b). In channels with a successively increased serration period, continuous single-cell observation showed that, in qualitative agreement with model predictions, cells slowed from 6 to 12 µm, presented even slower and saltatory movement in the 24-µm stretches and were unable to enter the smooth channel (Fig. 4c, d, Supplementary Video 8). The model also predicts a linear dependence of the propulsion force on actin flow speed. Actin flow has two mechanical components: polymerization pushes filaments from the front to the back, while actomyosin contraction pulls at the back[12,15–17]. Accordingly, inhibition of myosin II slowed down but did not stall cells under adhesive conditions (Extended Data Fig. 5a–f). Following dual inhibition of adhesion and contractility, cells still performed better in the serrated than in the smooth channels (Extended Data Fig. 5d–f), suggesting a direct relation between actin flow speed and locomotive force, as predicted. Next, we quantitatively imaged the actin flow of talin-KO and EDTA-treated cells in textured channels and found an increase in the actin retrograde flow relative to the substrate whenever the cells slowed between the topographical features (Fig. 4e–g, Extended Data Fig. 5g, Supplementary Video 9). This suggests that, in analogy with the molecular clutch formed by transmembrane receptors, topographical features allow the coupling of retrograde actin flow to a substrate. Finally, we devised an entirely orthogonal approach, in which we placed primary mouse T lymphocytes on surfaces that were either passivated or coated with the integrin ligand intracellular adhesion molecule 1 (ICAM1) and overlaid them with a pad of agarose. As shown previously[8], cells migrated on ICAM1, but slipped and did not migrate on passivated substrate (Extended Data Fig. 6a–g). We then used passivated surfaces with engraved linear ridges to introduce anisotropic topography. We found that on these surfaces, migratory capacity was rescued and that cells predominantly migrated perpendicular but not parallel to the ridges (Extended Data Fig. 6a–g, Supplementary Video 10). Within the same cell, actin showed retrograde flow in relation to the substrate when the flow was aligned with the ridges, but not in orientation perpendicular to the ridges (Extended Data Fig. 7, Supplementary Video 10). Together, these data demonstrate that the cells slipped in an orientation along the ridges and coupled in an orientation perpendicular to the ridges.

In our experiments, we demonstrated topography-based force transmission by controlling the topography of the environment, meaning that we extrinsically imposed shape changes of the cell. To directly link this to the amoeboid (shape-changing) principle, we tested whether cells can autonomously generate appropriate deformations. Morphometric analysis of T cells showed that they have the intrinsic capacity to produce rearwards-propagating deformation waves (Extended Data Fig. 8a–e and previously published studies[18–23]) and that the travelling speed of these waves scaled with the speed of actin flow. Such deformation waves could intercalate with any textured environment and propel a cell as conjectured by Abercrombie[2].

Altogether, our findings demonstrate that cells can transmit forces by coupling the retrograde flow of actin to a geometrically irregular environment, and that this can happen in the complete absence of any transmembrane receptors that link the cytoskeleton to the substrate. Notably, force transmission by specific adhesion receptors, by unspecific friction and by shape change are not mutually exclusive mechanisms. Rather, they are variants of the same fundamental principle of coupling

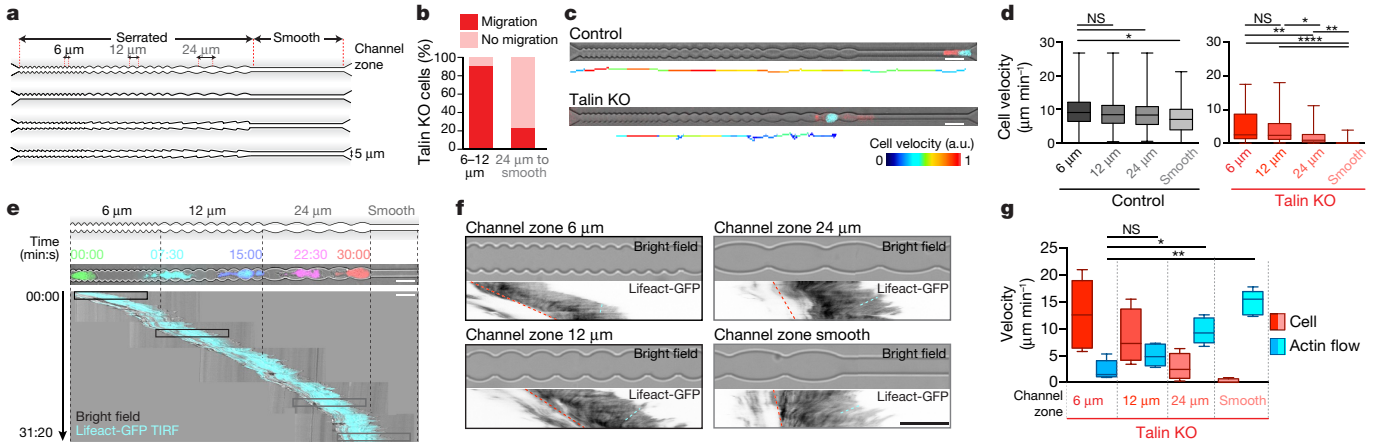


Fig. 4 | Force transmission of T cells in varying geometries. a, The channel design that is used in **b**–**g**. **b**, The proportion of talin-KO cells migrating in 6-µm and 12-µm zones ($n = 41$) and in the 24-µm zone and the smooth zone ($n = 22$). ****$P < 0.0001$, Fisher's exact test. **c**, Snapshots of control cells (top) and talin-KO cells in channels with a varying period. Lifeact-GFP is in red, and Hoechst is in cyan. Scale bars, 20 µm. Bottom, tracks are colour-coded for cell velocity. Data are representative of four and eight independent experiments, respectively. **d**, The velocity of control ($n = 88$) and talin-KO ($n = 79$) cells in zones with different periodicity. Data are representative of four and eight independent experiments, respectively. Control: *$P = 0.0234$; KO: *$P = 0.0188$, **$P = 0.0013$ (6 µm versus 24 µm) and **$P = 0.001$ (24 µm versus smooth), **$P < 0.0001$, otherwise NS, Kruskal–Wallis test followed by post hoc Dunn's test. **e**, TIRF microscopy of a Lifeact-GFP-expressing talin-KO cell moving from the 6-µm zone to the smooth zone. From top to bottom: scheme of the channel, colour-coded snapshots every 7.5 min and a kymograph of Lifeact-GFP (TIRF is shown in cyan and bright field is shown in grey). Scale bars, 20 µm. Representative of three independent experiments. **f**, Bright-field imaging of channel sectors and the corresponding kymograph of the Lifeact-GFP-expressing talin-KO cell shown in **e** during 50 s. The red dashed lines indicate cell displacement, and the cyan dashed lines indicate actin retrograde flow. Scale bar, 20 µm. **g**, Actin retrograde flow and cell velocities of talin-KO cells in different channel zones. Data are representative of four independent experiments. $n = 7$. *$P = 0.0423$ and **$P = 0.0036$, Kruskal–Wallis test followed by post hoc Dunn's test. Boxes in **d** and **g** extend from the 25th to the 75th percentile, with the middle line showing the median and the whiskers indicating the minimum to maximum values.

retrograde actin flow to the surrounding and can therefore operate alternately or even simultaneously. This endows amoeboid cells with enormous adaptability when migrating in heterogeneous environments.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2283-z.

1. Lämmermann, T. & Sixt, M. Mechanical modes of 'amoeboid' cell migration. *Curr. Opin. Cell Biol.* **21**, 636–644 (2009).
2. Abercrombie, M., Heaysman, J. E. & Pegrum, S. M. The locomotion of fibroblasts in culture. 3. Movements of particles on the dorsal surface of the leading lamella. *Exp. Cell Res.* **62**, 389–398 (1970).
3. Liu, Y.-J. J. et al. Confinement and low adhesion induce fast amoeboid migration of slow mesenchymal cells. *Cell* **160**, 659–672 (2015).
4. Friedl, P. & Wolf, K. Plasticity of cell migration: a multiscale tuning model. *J. Cell Biol.* **188**, 11–19 (2010).
5. Lämmermann, T. et al. Rapid leukocyte migration by integrin-independent flowing and squeezing. *Nature* **453**, 51–55 (2008).
6. Bergert, M. et al. Force transmission during adhesion-independent migration. *Nat. Cell Biol.* **17**, 524–529 (2015).
7. Schmidt, S. & Friedl, P. Interstitial cell migration: integrin-dependent and alternative adhesion mechanisms. *Cell Tissue Res.* **339**, 83–92 (2010).
8. Hons, M. et al. Chemokines and integrins independently tune actin flow and substrate friction during intranodal migration of T cells. *Nat. Immunol.* **19**, 606–616 (2018).
9. Woolf, E. et al. Lymph node chemokines promote sustained T lymphocyte motility without triggering stable integrin adhesiveness in the absence of shear forces. *Nat. Immunol.* **8**, 1076–1085 (2007).
10. Calderwood, D. A., Campbell, I. D. & Critchley, D. R. Talins and kindlins: partners in integrin-mediated adhesion. *Nat. Rev. Mol. Cell Biol.* **14**, 503–517 (2013).
11. Le Berre, M., Aubertin, J. & Piel, M. Fine control of nuclear confinement identifies a threshold deformation leading to lamina rupture and induction of specific genes. *Integr. Biol. (Camb.)* **4**, 1406–1414 (2012).
12. Renkawitz, J. et al. Adaptive force transmission in amoeboid cell migration. *Nat. Cell Biol.* **11**, 1438–1443 (2009).
13. Renkawitz, J., Reversat, A., Leithner, A., Merrin, J. & Sixt, M. Micro-engineered "pillar forests" to study cell migration in complex but controlled 3D environments. *Methods Cell Biol.* **147**, 79–91 (2018).
14. Vargas, P., Barbier, L., Sáez, P. J. & Piel, M. Mechanisms for fast cell migration in complex environments. *Curr. Opin. Cell Biol.* **48**, 72–78 (2017).
15. Medeiros, N. A., Burnette, D. T. & Forscher, P. Myosin II functions in actin-bundle turnover in neuronal growth cones. *Nat. Cell Biol.* **8**, 215–226 (2006).
16. Henson, J. H. et al. Two components of actin-based retrograde flow in sea urchin coelomocytes. *Mol. Biol. Cell* **10**, 4075–4090 (1999).
17. Cramer, L. P. Molecular mechanism of actin-dependent retrograde flow in lamellipodia of motile cells. *Front. Biosci.* **2**, d260–d270 (1997).
18. Driscoll, M. K. et al. Cell shape dynamics: from waves to migration. *PLoS Comput. Biol.* **8**, e1002392 (2012).
19. Howe, J. D., Barry, N. P. & Bretscher, M. S. How do amoebae swim and crawl? *PLoS One* **8**, e74382 (2013).
20. Bae, A. J. & Bodenschatz, E. On the swimming of *Dictyostelium* amoebae. *Proc. Natl Acad. Sci. USA* **107**, E165–E166 (2010).
21. Mandeville, J. T. H., Lawson, M. A. & Maxfield, F. R. Dynamic imaging of neutrophil migration in three dimensions: mechanical interactions between cells and matrix. *J. Leukoc. Biol.* **61**, 188–200 (1997).
22. Tozluoğlu, M. et al. Matrix geometry determines optimal cancer cell migration strategy and modulates response to interventions. *Nat. Cell Biol.* **15**, 751–762 (2013).
23. Jankowiak, G., Peurichard, D., Reversat, A., Schmeiser, C. & Sixt, M. Modelling adhesion-independent cell migration. *Math. Model. Methods Appl. Sci.* **30**, 513–537 (2020).

Methods

Cell culture

Cells were grown and maintained in a humidified incubator at 37 °C and 5% CO_2. LMR7.5 T cell hybridomas, a gift from A. M. Lennon-Duménil (Institut Curie, Paris, France), were cultivated in R10 medium (RPMI 1640 supplemented with 10% FBS, 2 mM L-glutamine, 100 units/ml penicillin, 100 µg/ml streptomycin and 50 µM 2-mercaptoethanol; Gibco, Thermo Fisher Scientific). The Lifeact-GFP[24] parental T cell line was generated by nucleofection with an eGFP reporter construct (Kit V, Lonza). Alternatively, LMR T cells were infected with a plasmid encoding Lifeact-mCherry (pLenti6.3, Invitrogen, Thermo Fisher Scientific). LX-293 HEK cells (Clontech) were used for lentivirus production and maintained in D10 medium (DMEM with 2 mM L-glutamine, 10% FBS, and 100 units/ml penicillin and 100 µg/ml streptomycin; Gibco). All cell lines were mycoplasma-free and frequently tested. Neutrophil-like PLB-985 cells were obtained from the DSMZ (ACC 139). In brief, cells were maintained in RPMI 1640, supplemented with 10% FBS, 20 mM HEPES and 1% glutamine (all Gibco, Thermo Fisher Scientific). For differentiation, PLB cells were kept for 6 days in medium containing 1.25% DMSO (cell culture grade; Sigma-Aldrich). Differentiation status was validated using flow cytometry (CD11b, Miltenyi Biotec). Before experimentation, differentiated cells were washed three times to remove DMSO. Dendritic cells were differentiated from bone marrow that originated from 6–12-week-old male or female C57BL/6J mice[25]. Primary mouse T cells were obtained from peripheral lymph nodes and spleens[8] and homogenized with a 70-µm cell strainer. Untouched primary naive T cells were isolated from 6–12-week-old male C57BL/6J Lifeact-GFP mice with an EasySep Mouse T cell Isolation Kit according to the manufacturer's protocol (19851A, STEMCELL Technologies) and incubated overnight at 37 °C in 5% CO_2 in R10 medium.

Animals

Mice were bred and maintained at the local animal facility in accordance with the IST Austria ethics commission and the Austrian law. Permission was granted by the Austrian Federal Ministry of Science, Research and Economy (identification code: BMWF-66.018/0005-II/3b/2012).

Reporter and CRISPR–Cas9 cell line generation

CRISPR–Cas9 design and production. Single-guide RNAs (sgRNAs) were designed as described elsewhere[26]. In brief, sgRNAs were designed to induce the double-strand break in the first exons of the mouse gene encoding talin 1 (*Tln1*). Then sgRNAs were scored for high on-target effects (https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design) and low off-target effects (http://crispr.mit.edu). sgRNAs were cloned into the LentiCRISPRv1 vector (a gift from F. Zhang, Addgene plasmid no. 49535). Guide sequences: sgRNA-Scramble s(5′–3′) GCCGTGGCGCATGGGTAGCA; sgRNA-Talin1g1 s(5′–3′) ATAATGCCCTAC GAGCCGT (off-target score of 91); sgRNA-Talin1g2 s(5′–3′) CTCACT GTTTCCCCGGGTA (off-target score of 82); sgRNA-Talin1g3 s(5′–3′) GTCGAGGCTGGGCGACTGG (off-target score of 98).

Lentivirus production. Lentivirus production was performed as described previously[25]. In brief, LX-293 HEK cells (Clontech) were co-transfected with LentiCRISPRv1 or LentiCRISPRv2, packaging- psPAX2 (Addgene no. 12260) and pCMV-VSV-G envelope plasmids using Lipofectamine 2000 (Thermo Fisher Scientific) as recommended by the manufacturer, and cells were resuspended in R10 medium 1 day before transfection. Supernatant was collected after 72 h and stored at −80 °C.

CRISPR-based talin KO. Lifeact-eGFP LMR7.5 T cells were spin-infected at 1,500*g* for 1 h in the presence of the lentivirus-containing supernatant and 6 µg/ml Polybrene (Sigma-Aldrich). Spin infection was carried out in 12-well plates with 3×10^5 cells/ml per reaction plus 500 µl of undiluted virus. Three days after infection, cells were selected for stable virus insertion using 5 µg/ml puromycin (Gibco, Thermo Fisher Scientific) for 7 days and were subjected to western blotting. Cells were then cloned using fluorescence-activated cell sorting (FACS Aria III, BD Biosciences) and expanded in R10 medium containing puromycin (5 µg/ml).

T cell infection and selection. For infection of LMR7.5 T cells with the Lifeact-mCherry reporter construct, the same infection protocol was applied except cells were selected with 5 µg/ml blasticidin (Gibco, Thermo Fisher Scientific) and sorted for fluorescence.

Western blot

For immunodetection of talin and GAPDH proteins, 1×10^6 control and talin-KO cells were centrifuged and lysed in RIPA buffer (New England Biolabs) and the supernatant was denatured in 2× Laemmli buffer (161-0737, Bio-Rad,) at 95 °C for 5 min. Protein lysates were loaded on a 10% Tris-glycine gel (Invitrogen, Thermo Fisher Scientific), separated by SDS–PAGE and further transferred by electrophoresis to a polyvinylidene difluoride membrane (iBlot, Invitrogen, Thermo Fisher Scientific) according to the manufacturer's instructions, with a 9-min transfer time. Membranes were blocked in PBST-BSA 5% for 1 h and incubated with primary antibodies overnight at 4 °C. Antibodies to talin (mouse anti-pan-talin, clone 8d4; T3287, Sigma-Aldrich) and GAPDH (mouse, clone GA1R; ab125247, Abcam) were diluted 1:200 and 1:3,000 in PBST-BSA 1%, respectively. After washing with TBST, secondary antibody was applied for 1 h in PBST-BSA 5% at room temperature (goat anti-mouse IgG HRP conjugate, diluted 1:5,000; 170-6516, Bio-Rad). Protein detection was performed by enhanced chemiluminescence (ECL; Thermo Fisher Scientific) detection using a VersaDoc imaging system (Bio-Rad).

Adhesion assay

Control and talin-KO T cells (5×10^5) were plated in pre-warmed R10 medium in a 25-cm^2 culture flask (TPP, Sigma-Aldrich) and placed in a humidified incubator at 37 °C, 5% CO_2 for 24 h before observation. Cells were imaged with an inverted microscope at low magnification (DM IL Led and DFC450 digital camera, Leica Microsystems) and cells adhering to the bottom of the dish were counted manually.

Spreading assay

The 2D spreading assay was performed on glass coverslips (Menzel-Glaser no. 1, VWR) that were previously coated with 2 µg/ml of recombinant mouse ICAM-1/human Fc chimaera (796-IC, R&D Systems) overnight at 4 °C, and washed three times with PBS. Control (Lifeact-mCherry or Lifeact-GFP) and talin-KO (Lifeact-GFP or Lifeact-mCherry) T cells (0.2×10^6) were gently mixed in 200 µl R10 medium and allowed to settle on the coated dish for 60 min at 37 °C and 4.5% CO_2. Cell spreading was then imaged by simultaneous TIRF and epifluorescence microscopy.

Photolithography

Photolithography was performed as previously described[13,25,27,28] with the following adaptations.

Photomasks. Patterns were designed with Coreldraw X8 (Corel Corporation) exported to DXF, then converted to GERBER format with LinkCad. Chrome photomasks (100 mm or 125 mm; PhotoData/JD Photo-Tools) were used.

Confiners. The 3-µm confiner masters were produced by spin-coating SU8-GM1050 (Gersteltec) for 40 s at 3,000 rpm, pre-baking for 1 min at 120 °C, exposing to 35 mJ/cm^2 of UV, post-exposure baking for 5 min at 95 °C, developing in SU8 developer, and then hard-baking at 150 °C for 5 min. The 5-µm confiner master was produced by spin-coating SU8-2005 (Microchem) for 30 s at 3,000 rpm, pre-baking for 2 min at 95 °C, exposing to 105 mJ/cm^2 of UV, post-exposure baking for 3 min at 95 °C, developing in SU8 developer, and then hard-baking at 150 °C for 5 min.

EDTA device. A 4-µm high master was produced by spin-coating SU8-2005 for 30 s at 5,000 rpm, pre-baking for 2 min at 95 °C and UV exposure of 550 mJ/cm^2 through a PL-360-LP (Omega Optical) optical filter on an EVG mask aligner 610 (EVG Group). The wafer was post-exposure baked for 3 min at 95 °C, developed in SU8 developer and then hard-baked at 150 °C for 5 min. The optical filter was necessary for high-resolution features in SU8.

Multilayer devices. First, the 25-µm high control master was made by spin-coating SU8-3025 (Microchem) for 30 s at 3,000 rpm. The wafer was soft-baked for 15 min at 95 °C, then exposed to 200 mJ/cm^2 UV, then post-exposure baked at 95 °C for 5 min, then developed in SU8 developer. Second, the fluid-layer master was first fully coated with a submicron layer of GM1040-SU8 by spin-coating at 5,000 rpm for 2 min. The wafer was then baked at 95 °C for 5 min, exposed to 100 mJ/cm^2 UV, post-exposure baked at 95 °C for 30 min, and then developed in SU8 developer. The 5-µm fluid-layer features were made on top of the coated wafer by spin-coating SU-2005 for 30 s at 3,000 rpm, soft-baking for 2 min, UV exposure of 550 mJ/cm^2 through the optical filter, post-exposure baking for 3 min at 95 °C, and developing in SU8 developer. Third, the fluid-layer master was then prepared for the next layer by spin-coating HMDS at 3,000 rpm for 30 s and baking for 1 min at 126 °C. Scotch Magic tape (3M) was applied over the alignment marks. AZ-40XT-11D (MicroResist Technologies) was then coated at 3,000 rpm for 30 s. The Scotch tape was then peeled off to reveal the alignment marks. The wafer was then baked at 126 °C for 7 min, aligned and exposed to UV at 500 mJ/cm^2. After a post-exposure bake for 10 min at 120 °C, it was developed in 726 MIF for about 5–10 min. The features were then rounded to make parabolic channels at 130 °C for 2 min (centre height of 26.6 µm).

Silanization. Before applying PDMS, wafers were placed in a desiccator with 10 µl of Trichloro(1H,1H,2H,2H-perfluorooctyl)-silane (Sigma-Aldrich), a vacuum of 100 mbar was applied, and then the desiccator was sealed for 1 h. One permanent coating was sufficient for later uses.

PDMS device microfabrication

Microfabricated devices were generated as previously described[11,13,28], with 1:10 PDMS and degassed for 2 min at 2,000 rpm (mix) and for 2 min at 2,200 rpm (defoam) in a mixer/defoamer (ARE-250, Thinky) before use.

Confiner. PDMS (Sylgard 184, Ellsworth Adhesives) was gently poured onto the wafers, then 10-mm coverslips were activated by plasma cleaning for 2 min at medium intensity (Harrick Plasma Cleaner, pdc-002, Harrick Plasma) and pressed upside-down onto the PDMS-covered wafers. The wafer was baked on a hot plate at 95 °C for 15 min, and the 300-µm width micropillar-coated coverslip was gently removed from the wafer. A soft PDMS 10-mm diameter pillar (1:30) was poured into a homemade metal mould, degassed under vacuum and baked at 80 °C for 1 week.

EDTA device. Of 1:10 PDMS (Sylgard 184, Ellsworth Adhesives), 20–25 g was poured onto the wafer contained in an aluminium mould in a Petri dish, degassed in a vacuum desiccator and baked overnight at 80 °C. The devices were then diced with a razor blade and 2-mm entry and exit holes were punched (Harris Unicore biopsy puncher, Sigma-Aldrich). Following punching, they were cleaned with tape, sonicated in ethanol, blown dry and plasma-bonded to a coverslip.

Multilayer devices. PDMS (80 g; RTV615, Techsil) was used; 10 g of PDMS was put on the control layer and spin-coated (500 rpm for 15 s followed by 2,300 rpm for 60 s), and 70 g was applied on top of the flow layer in a tight aluminium mould and further degassed in a vacuum desiccator to pump out the remaining bubbles. Both layers were baked at 80 °C for 45 min, and the flow layer was removed from the mould, trimmed with a razor blade, and entry and exit holes were punched with a homemade arbor press. Debris from both the flow device and the control layer on its wafer were carefully removed with Scotch tape, the two layers were plasma-cleaned (2 min at medium intensity), aligned manually and bonded together. After overnight baking in 80 °C, the chip was removed from the wafer, the control layer entry holes were punched, and the chip was bonded to the glass coverslip by oxygen plasma and kept at 80 °C until further use. Before the experiments, all entry ports were connected with metal pins (NE-13-1003, New England Small Tube Corporation).

Migration assays

Before experiments, 2×10^5 T cells in 2 ml R10 medium were stained with Hoechst 33342 for 30 min (1 drop, NucBlue, R37605, Invitrogen, Thermo Fisher Scientific) when nucleus visualization was needed for cell tracking.

Collagen assay. A 3D collagen scaffold (final concentration of 1.7 mg/ml) was obtained by mixing bovine collagen (PureCol, Advanced BioMatrix) in 1× minimum essential medium Eagle and 0.4% sodium bicarbonate (both Sigma-Aldrich), with 3×10^5 cells in R10 medium at a 2:1 ratio[5,29]. After casting the mix in homemade migration chambers, gels were allowed to polymerize for 45 min at 37 °C, 5% CO_2. Migration was observed by time-lapse video microscopy.

Cell confiner assay. PDMS micropillars were placed on the soft pillar on a homemade magnetic glass lid. When indicated, dish and confiner were coated with 2 µg/ml of recombinant mouse ICAM-1/human Fc chimaera (796-IC, R&D Systems). The setup was placed in a dish containing R10 medium in a humidified incubator at 37 °C, 5% CO_2, 1 h before the experiments. Cells (5×10^4 in 5 µl R10 medium) were then pipetted onto the medium-freed micropillars and confined on a dish above a magnetic ring, and R10 medium was added back in the dish ready for imaging. For the myosin inhibition experiment, 50 µM para-nitroblebbistatin[30] was added to the medium (Optopharma).

Primary T cell confinement assay. To measure actin flow speed and cell curvatures, we used an agarose-confinement setup[8]. In brief, plasma-activated glass was overlaid with a 0.2 mg/ml solution of poly(L-lysine)-graft-PEG copolymer (PLL(20)-g[3.5]-PEG(2); SuSoS) at 4 °C overnight. To form a 0.5% agarose block with 10% FBS (and standard concentrations of other supplements), one part 2× HBSS buffer (Sigma) and two parts RPMI supplemented with twice the concentrations of all other supplements used in R10 medium were mixed together with one part 1% high-molecular-weight agarose (850152, Biozym Scientific GmbH) in water at 50 °C and subsequently cast onto the dish then allowed to solidify at room temperature. CCL19 (250-27B, Peprotech) was added to soluble agarose before casting. T cells were injected under the agarose block with a micropipette and incubated for 30 min at 37 °C under 5% CO_2 before imaging.

EDTA migration assay. T cells or PLB cells (5×10^4) were introduced into the 2-mm entry port of the microchamber, and the device was soaked in R10 medium. Cells were left to enter freely into the confinement and channel zone for 1 h and then imaged with a video microscope. After 1 h of imaging, chemical disruption of the cell adhesions was performed by removing R10 medium from the dish and adding R10 medium containing 10 mM EDTA (prepared from 0.5M stock; EDS-100G, Sigma-Aldrich) with or without 50 µM para-nitroblebbistatin[30] (Optopharma). Imaging was then resumed for 1 h. For dendritic cells, experiments were performed 1 day after overnight LPS activation (200 ng/ml; Sigma-Aldrich) with or without

2.5 µg/ml CCL19 (250-27B, Peprotech) added on the opposite hole of cell entry and with or without 30 mM EDTA[13].

Multilayered microfluidics: maze and microchannels migration assay. As talin-KO cells failed to freely enter the confinement zone, a homemade microfluidic setup was used to push the cells by applying a pressure differential to the inlet and outlet ports of the device. The microfluidics designs in Extended Data Fig. 2d were used, as well as separated chambers containing the four pillar mazes (identical to ref. [28] with four chambers). In brief, the core component of those devices is the flow layer with a 5-µm height pillar maze (Fig. 2a–d, Extended Data Fig. 2e) or channels (Figs. 2e–g, 4, Extended Data Figs. 2f, 4) with an exit, cell and medium entry ports and on one side and on the other side a sink channel connected to two or three exits. Flow and sink channels are rounded, 26 µm high. For a controlled flow of cells and fluids, all ports are equipped with independently controllable push-up PDMS membrane valves. In brief, the control channels are connected to solenoid valves (MH1, miniature, Festo) controlled with a MATLAB graphical user interface (MathWorks). First, the chip was mounted into the 37 °C, 5% CO_2 chamber of the microscope, and optimal closing pressures of 0.15–0.2 MPa of the water-filler PDMS membrane valves were determined for each chip; each valve was checked individually by microscopy. Second, the whole chip was saturated with a pre-warmed medium by applying a pressure of 0.04 MPa and incubated in R10 medium for 30 min. Cells were first loaded via the ports on one side, resulting in their distribution along the chamber, then slowly pushed with a pressure of <0.01 MPa from the medium port towards the confinement zone, and further pushed in the microchannels or micropillars maze. Finally, all valves were closed to ensure that no external pressure or flow was applied to the cells during imaging.

TIRF microfluidics: microchannels migration assay. For Fig. 4e–g, TIRF microscopy was not possible with the multilayered microfluidic devices, so devices with the flow layer only were generated. First, R10 medium was forced into the device with a vacuum desiccator, and then the chips were incubated in humidified 5% CO_2 at 37 °C for 2 h before the experiments. Experiments were performed on the microscope stage (humidified environment 5% CO_2 at 37 °C) and a low cell flow was delivered and stopped after talin-KO cell loading in the microchannel using a LA120 syringe pump (5 µl/h; Landgraf Laborsysteme).

Primary T cell confinement assay on nanoridges

T cells were confined under agarose as previously described with minor modifications[8]. In brief, plasma-activated coverslips with nanoridges (800 nm/800 nm/600 nm (groove/ridge/depth)) (ANFS-CS25-50, NanoSurface Biomedical) were passivated with a 1 mg/ml solution of poly(L-lysine)-graft-PEG copolymer (PLL(20)-g[3.5]-PEG(2), SuSoS) at 4 °C overnight. Flat control coverslips (FLAT-CS25-50, NanoSurface Biomedical) were either passivated as described above or coated with rhICAM1-Fc (2 µg/ml) (720-IC-050, R&D) to generate adhesive substrates. Next, 0.5% agarose was prepared by mixing A, 5 ml of 2× HBSS buffer (Sigma), and mixing B, 10 ml of serum-free RPMI supplemented with 20% BSA and twice the concentrations of all other supplements used in R10 medium (see above), and mixing C, 5 ml of 1% high-molecular-weight agarose (850152, Biozym Scientific GmbH) in PBS. Agarose solution (0.5%) was kept at 50 °C in a water bath before it was layered on top of the coated coverslips and allowed to solidify for 1 h at 4 °C. CCL19 (10 nM; 250-27B, Peprotech) was added to soluble agarose before casting. T cells were isolated from the spleen of wild-type and/or Lifeact-GFP mice (T cell isolation kit mouse; MACS, Miltenyi Biotec) and kept in R10 medium at 37 °C and 5% CO_2. Before imaging (30 min), 0.5 µl of highly concentrated T cells was injected under the agarose block.

Time-lapse video microscopy

Bright-field video of T cells for collagen assays (×10 objective), 3-µm high confiner assays and primary T cells migrating under agarose (×20 objective) were acquired by time-lapse acquisition (time interval of 20 s) using inverted cell culture microscopes (DM IL Led, Leica Microsystems) equipped with cameras (ECO415MVGE, SVS-Vistek) and custom-built climate chambers (5% CO_2, 37 °C, humidified). A 5-µm high confiner, EDTA and microfluidics assays were recorded every 30 s or 60 s at 2–6 multi-positions with NIS Elements software (Nikon Instruments). Experiments were performed with an inverted wide-field Nikon Eclipse Ti microscope in a humidified and heated chamber at 37 °C and 5% CO_2 (Ibidi Gas Mixer), equipped with a ×20/0.5 NA PH1 air objective, a Hamamatsu EMCCD C9100 camera and a Lumencor Spectra X light source (390 nm, 475 nm, 542/575 nm; Lumencor). TIRF microscopy was performed with a ×60/1.46 NA oil objective, optovar ×1 or ×1.6 in a humidified and heated chamber at 37 °C and 5% CO_2 using an inverted Axio Observer (Zeiss) microscope, a TIRF 488/561-nm laser (Visitron Systems) and an Evolve EMCCD camera (Photometrics) controlled by VisiView software (Visitron Systems). To visualize actin retrograde flow, assays were recorded every second for TIRF acquisition only (Fig. 1e–g, Supplementary Video 3 example 3) or every 2 s or 3 s when acquisition of both TIRF or wide-field fluorescence and bright field was necessary (Fig. 4e–g, Extended Data Figs. 4b, 5g, Supplementary Videos 3 (examples 1 and 2), 10). When cells migrated outside the field of view (Supplementary Video 10), they were manually followed via a joystick controller. For the primary T cell confinement assay on nanoridges (Extended Data Fig. 7), a video of actin flow (Lifeact-GFP) were recorded (2-s frame rate) on an inverted spinning-disc confocal microscope (Andor) using a ×100/1.4 NA objective and a 488-nm laser line in a custom-built climate chamber (37 °C under 5% CO_2). For the primary T cell flow analysis (Extended Data Fig. 8), inverted spinning-disc confocal microscopy was performed at 37 °C under 5% CO2 with an iMIC (TILL Photonics, FEI) instrument with a ×60/1.35 NA objective and a 488-nm laser line.

Image analysis

FIJI imaging processing software (https://fiji.sc/) was used for image and video microscopy analysis.

Spreading area. TIRF images of single cells were binarized after de-speckling and background subtraction, and the spreading area was measured using the Analyse Particles tool.

Cell size measurement. To measure cell size, fluorescent images of 3-µm-confined control (Lifeact-mCherry) and talin-KO (Lifeact-GFP) cells were binarized, and the single-cell spreading area and perimeter were measured using the Analyse Particles tool.

Manual tracking. A 1-h-long bright-field video of T cells in collagen and 3-µm high confiners were reduced to a 1-min and 2-min interval, respectively, and cells were tracked manually by using the 'Manual tracking' plugin provided by Fiji.

Automated tracking. For the 5-µm high confiner, pillar maze and EDTA assays, cell migration was analysed by nucleus tracking using TrackMate[31] (https://imagej.net/TrackMate). For the EDTA assay, tracks before and after EDTA treatment were analysed and the speed was extracted at the single-cell level in the different zones (2.5D confinement, smooth and serrated microchannels containing only one cell). For Fig. 4b, cell migration in the channel zones was regarded as efficient for velocities above 0.5 µm/min. The mean square displacement was calculated from the tracking files with a homemade script (MATLAB, MathWorks).

Actin flow versus cell velocity analysis. For Figs. 1e–g, 4e–g, Extended Data Figs. 1l, 4b, 5g, 8a, kymographs were generated along the cell trajectories, and both cell and actin velocities were extracted from those kymographs. To follow cells migrating in a channel with a high magnification (Fig. 4e–g, Supplementary Video 10), the channel was realigned with a homemade script (MATLAB, MathWorks).

Microchannel assay analysis. Cells going in both directions of the microchannels were analysed, and channels containing more than one cell were excluded. Kymographs were generated along the channel, and cell velocity was calculated for each cell in the different zones.

Actin flow in primary T cells confined on nanoridges. Actin flow vectors were measured by two methods: A, which was automatic particle tracking with TrackMate plugin (3.8.0) of FIJI[31] (ImageJ 1.52p, java 1.8.0), and B, which was optical flow analysis (using a custom MATLAB code). Only frames where optical flow analysis was confirming automatic particle tracking (deviation of <30°) were subjected to further analysis. To estimate vectors of forwards cell movement, cell contours were segmented with Ilastik[32], and cell displacement was measured by tracking the centroids. Angular distributions of actin flow vectors and angular distributions of cell forwards movement vectors were then analysed and plotted with respect to the orientation of nanoridges using a custom MATLAB code.

Curvature analysis. For Extended Data Fig. 8, we set up a homemade script that performs binarization of fluorescent images of Lifeact-GFP-expressing cells. The local curvature was calculated from a spline fit to the cell outline, and the cortical actin flow was quantified with optical flow analysis (MATLAB, MathWorks).

Statistical analysis

All statistical analyses were performed with GraphPad Prism (GraphPad Software). The normality test was performed, and the statistical strategy used accordingly (for example, non-parametric Mann–Whitney *U*-test and one-way ANOVA with Kruskal–Wallis test with post hoc Dunn's test) as indicated in Supplementary Table 1. For the box and whiskers graphs in Fig. 4d, g and Extended Data Figs. 2f, 4a, 4b, 5g, 6f, the box extends from the 25th to the 75th percentiles, with the middle line showing the median, and the whiskers indicating the minimum to maximum values. See Supplementary Table 1 for additional information.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data that support the findings of this study are available within the Article and its Supplementary Information. Source data for Figs. 1, 2, 4 and Extended Data Figs. 1–8 are provided with the paper. Original videos will be made available on reasonable request due to the large file size.

24. Riedl, J. et al. Lifeact: a versatile marker to visualize F-actin. *Nat. Methods* **5**, 605–607 (2008).
25. Leithner, A. et al. Diversified actin protrusions promote environmental exploration but are dispensable for locomotion of leukocytes. *Nat. Cell Biol.* **18**, 1253–1259 (2016).
26. Shalem, O. et al. Genome-scale CRISPR–Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
27. Leithner, A., Merrin, J., Reversat, A. & Sixt, M. Geometrically complex microfluidic devices for the study of cell migration. *Protoc. Exch.* https://doi.org/10.1038/protex.2016.063 (2016).
28. Schwarz, J. et al. A microfluidic device for measuring cell migration towards substrate-bound and soluble chemokine gradients. *Sci. Rep.* **6**, 36440 (2016).
29. Sixt, M. & Lämmermann, T. in *Cell Migration: Developmental Methods and Protocols* (eds. Wells, C. M. & Parsons, M.) 149–165 (Humana, 2011).
30. Képiró, M. et al. *para*-Nitroblebbistatin, the non-cytotoxic and photostable myosin II inhibitor. *Angew. Chem. Int. Ed. Engl.* **53**, 8211–8215 (2014).
31. Tinevez, J. Y. et al. TrackMate: an open and extensible platform for single-particle tracking. *Methods* **115**, 80–90 (2017).
32. Sommer, C., Straehle, C., Koethe, U. & Hamprecht, F. A. Ilastik: interactive learning and segmentation toolkit. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 230–233 (IEEE, 2011).

Acknowledgements We thank A. Leithner and J. Renkawitz for discussion and critical reading of the manuscript; J. Schwarz and M. Mehling for establishing the microfluidic setups; the Bioimaging Facility of IST Austria for excellent support, as well as the Life Science Facility and the Miba Machine Shop of IST Austria; and F. N. Arslan, L. E. Burnett and L. Li for their work during their rotation in the IST PhD programme. This work was supported by the European Research Council (ERC StG 281556 and CoG 724373) to M.S. and grants from the Austrian Science Fund (FWF P29911) and the WWTF to M.S. M.H. was supported by the European Regional Development Fund Project (CZ.02.1.01/0.0/0.0/15_003/0000476). F.G. received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 747687.

Author contributions A.R. and M.S. conceived the experiments and wrote the manuscript, with critical feedback from all authors. A.R. designed, performed and analysed the experiments, with the help of I.d.V., J.S., M.H., R.H. and S.T. A.R. and J.M. designed the microfluidic devices. J.M. performed the photolithography. F.G. designed and performed the nanoridge experiments. J.A. fabricated the nanoridge substrates. R.H. wrote the image analysis scripts. A.C.-J., R.V., J.M., R.H., M.S. and A.R. discussed the physical model. A.C.-J. and R.V. wrote the physical model.

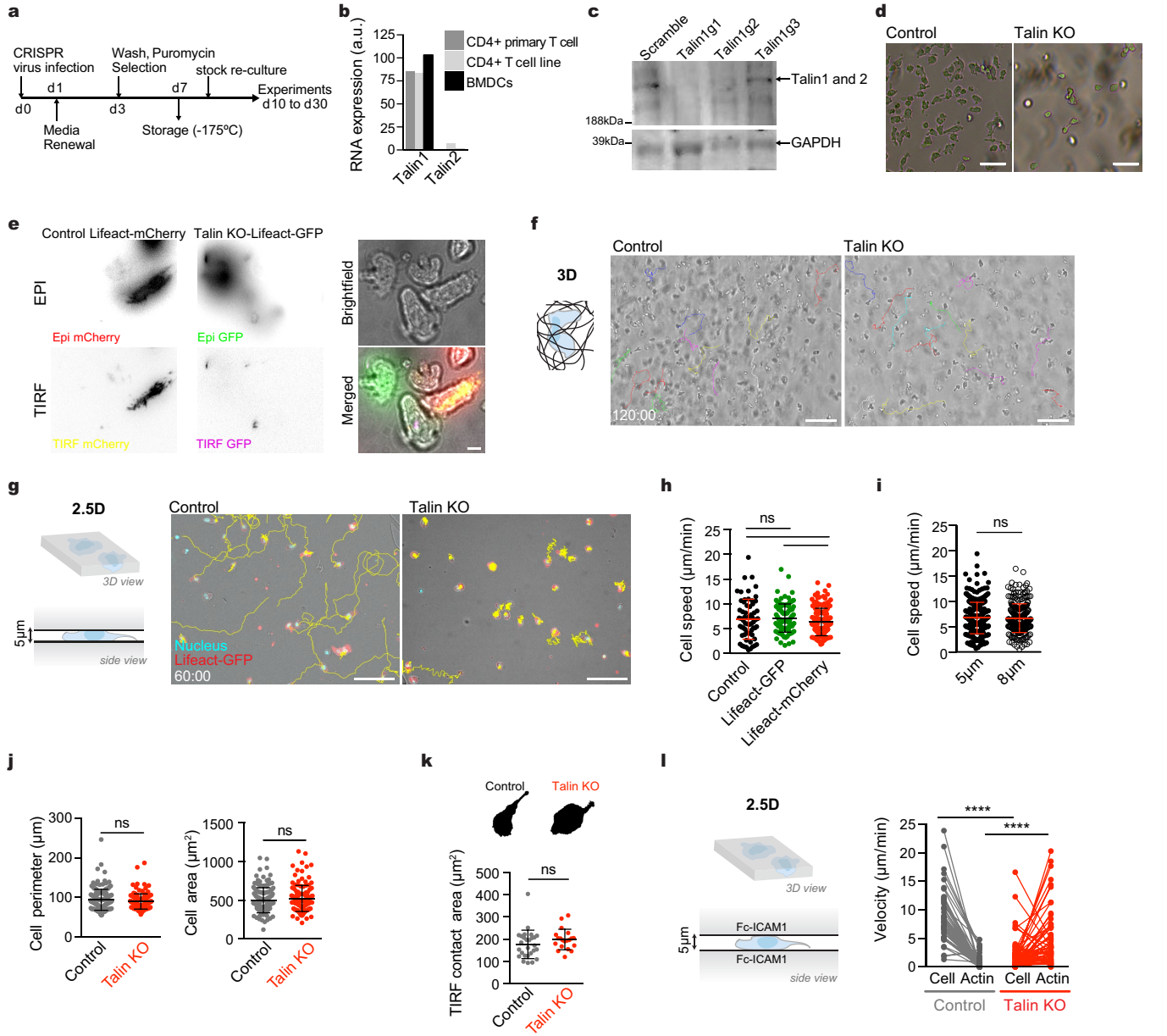
Competing interests The authors declare no competing interests.

Additional information

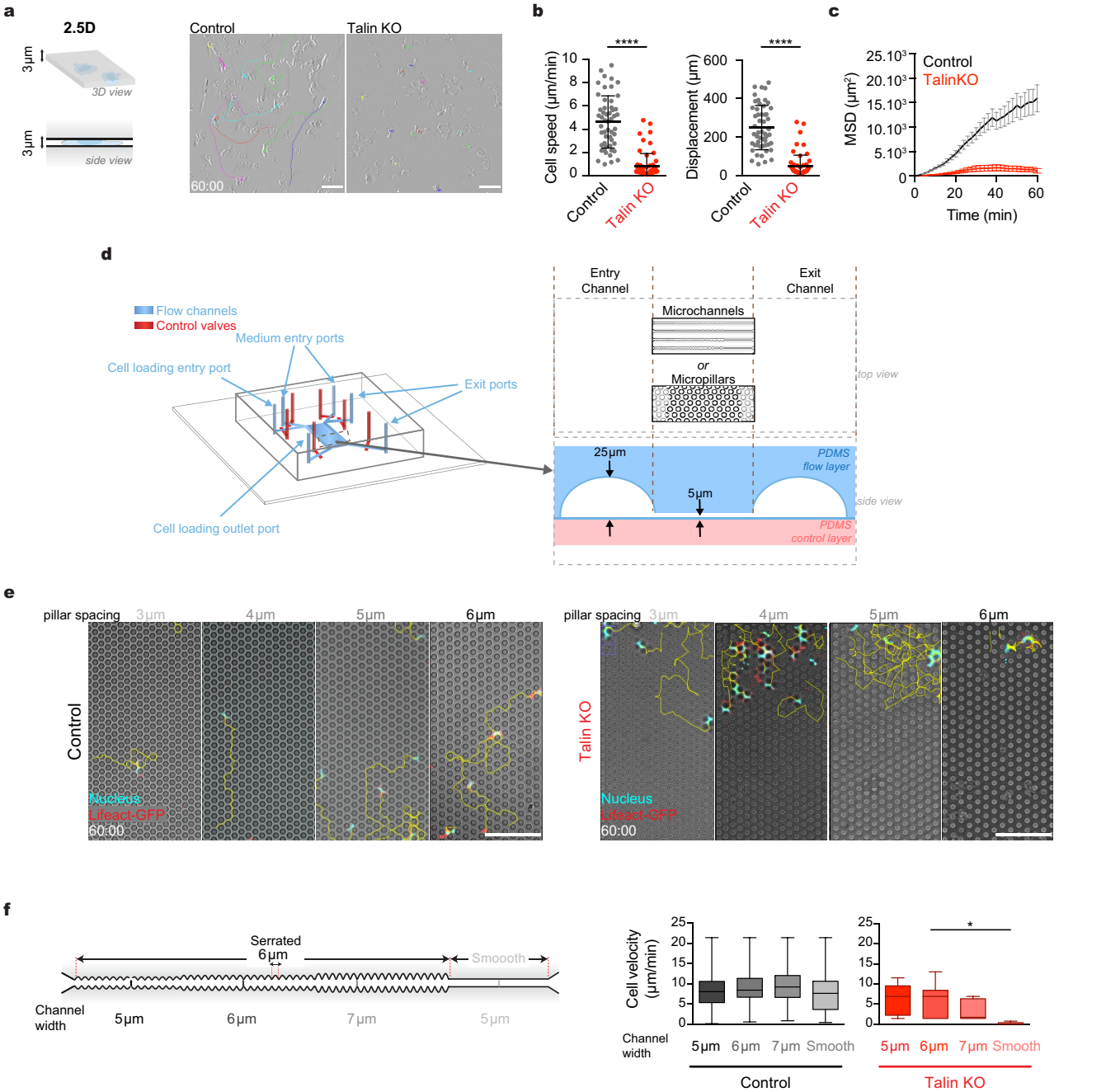
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2283-z.

Correspondence and requests for materials should be addressed to A.R. or M.S.

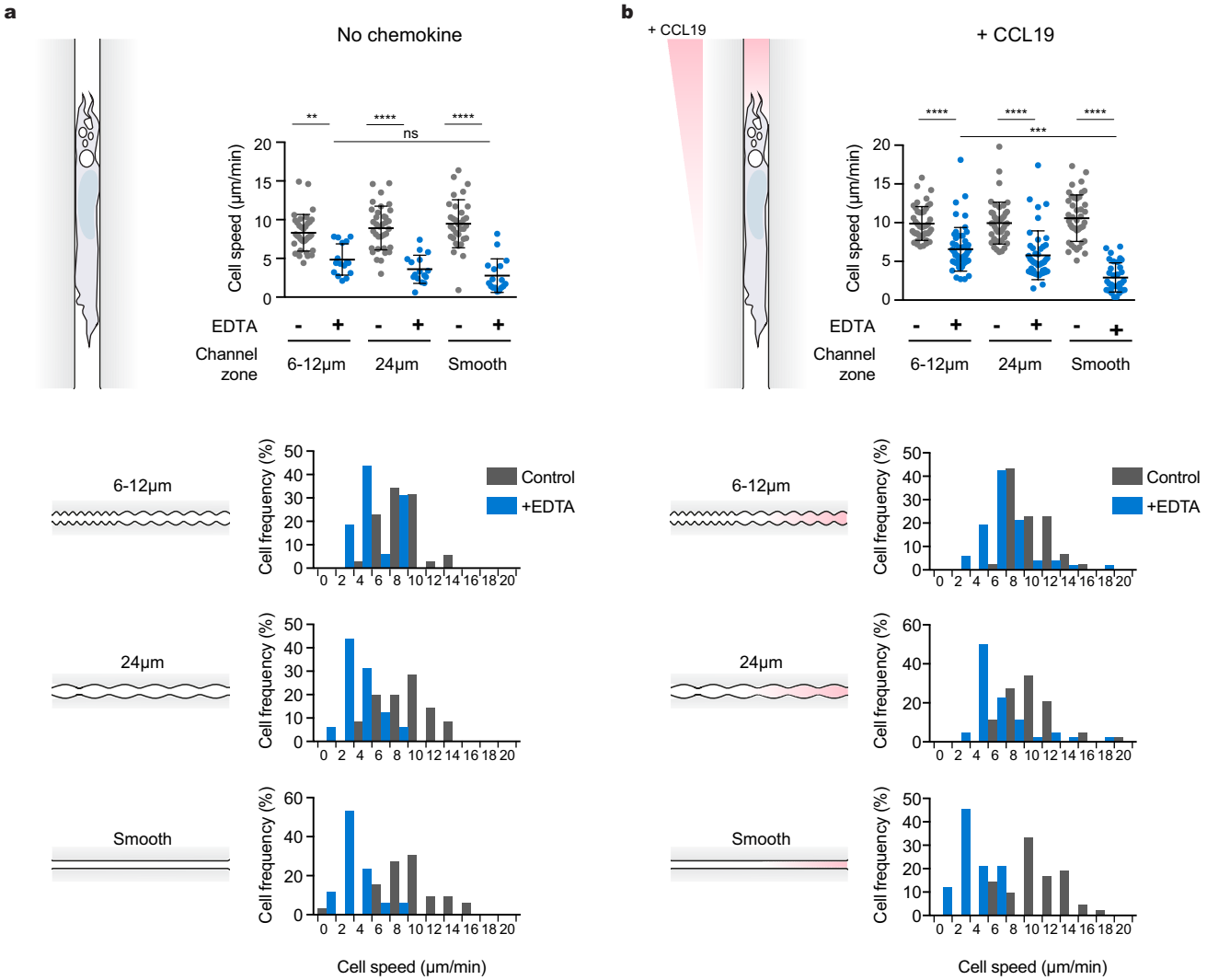
Peer review information *Nature* thanks Kenneth Yamada and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



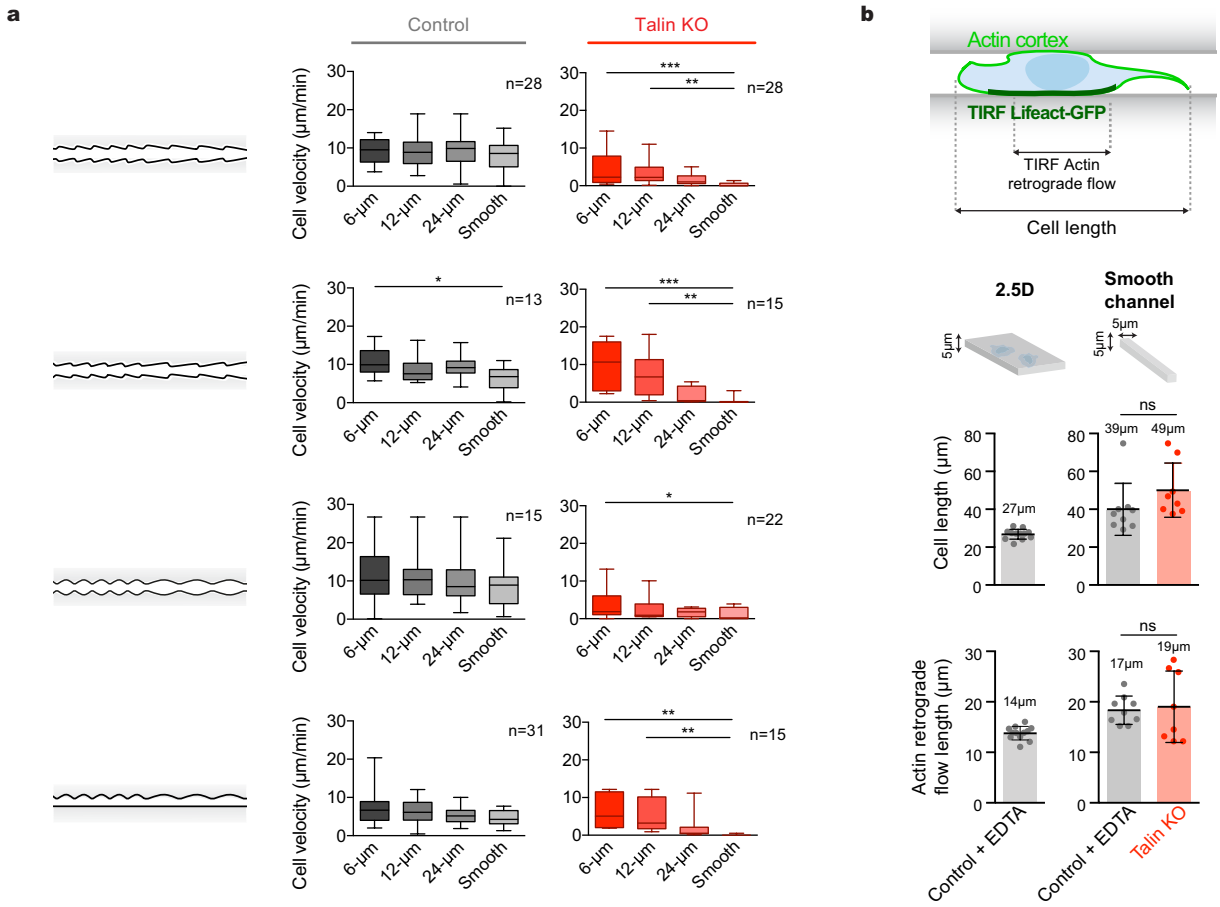
Extended Data Fig. 1 | Effects of *Tln1* deletion on T cell adhesion and migration. a, Workflow scheme of virus production, cell infection and analysis. d, day. **b**, RNA-seq analysis of talin 1 and 2 expression in mouse CD4$^+$ primary T cells, the T cell line and bone marrow-derived dendritic cells (BMDCs). **c**, Western blot analysis of talin 1 and talin 2 expression in T cells infected with lentivirus containing scrambled and talin 1-targeting CRISPR guides (guides 1–3 (Talin1g1–Talin1g3)). Data are representative of three independent experiments. **d**, Related to Fig. 1a. Phase-contrast images of control and talin-KO cells. Scale bars, 50 µm. Representative of four experiments. **e**, Related to Fig. 1b. Representative snapshots of epifluorescence (EPI) and TIRF microscopy of control cells (left) and talin-KO cells (middle) expressing mCherry and eGFP Lifeact fusions, respectively, plated on a Fc-ICAM1-coated surface. Representative of three experiments. The right panel shows bright field and merged fluorescence. Scale bar, 5 µm. **f**, Related to Fig. 1c. Representative snapshots of control and talin-KO cells embedded in 3D collagen gel at $t = 120$ min; individual tracks are displayed in different colours. Representative of three experiments. Scale bars, 100 µm. **g**, Related to Fig. 1d. Left, scheme of 5-µm high confinement used in **f**, **g** and **j**. Right, control cells (left) and talin-KO T cells (right) expressing Lifeact-GFP (red), stained with Hoechst (cyan), were placed under 5-µm height confinement and imaged by video microscopy. Representative of four experiments. Snapshot is at $t = 60$ min. Nucleus tracks are displayed in yellow. Scale bars, 50 µm. **h**, Cell speed of control T cells ($n = 62$) and T cells expressing Lifeact-GFP ($n = 100$) or Lifeact-mCherry ($n = 138$). Data are mean ± s.d.; ns $P = 0.0775$, Kruskal–Wallis one-way ANOVA followed by two-sided post hoc Dunn's test; control versus Lifeact-GFP or Lifeact-mCherry: $P > 0.999$; Lifeact-GFP versus Lifeact-mCherry: $P = 0.0717$. **i**, Speed of T cells placed under 5-µm ($n = 300$) or 8-µm ($n = 274$) confinement. Data are mean ± s.d.; ns $P = 0.8460$, two-sided Mann–Whitney *U*-test. **j**, Perimeter (left) and area (right) of control and talin-KO cells placed under 3-µm confinement (control cells: $n = 145$, talin-KO cells: $n = 127$). Representative of three experiments. Data are mean ± s.d.; ns $P = 0.4261$ for perimeter and $P = 0.2558$ for area, two-sided Mann–Whitney *U*-test. **k**, Spreading area observed by TIRF microscopy of control ($n = 27$) and talin-KO ($n = 20$) cells under 5-µm confinement. Representative of two experiments. Data are mean ± s.d.; ns $P = 0.2025$, two-sided Mann–Whitney *U*-test. **l**, Scheme of the 5-µm high Fc-ICAM1-coated confinement (left) used to measure velocities and F-actin retrograde flow (right) in control cells ($n = 49$) and talin-KO T cells ($n = 59$). Representative of five experiments. ****$P < 0.0001$, two-sided Mann–Whitney *U*-test.



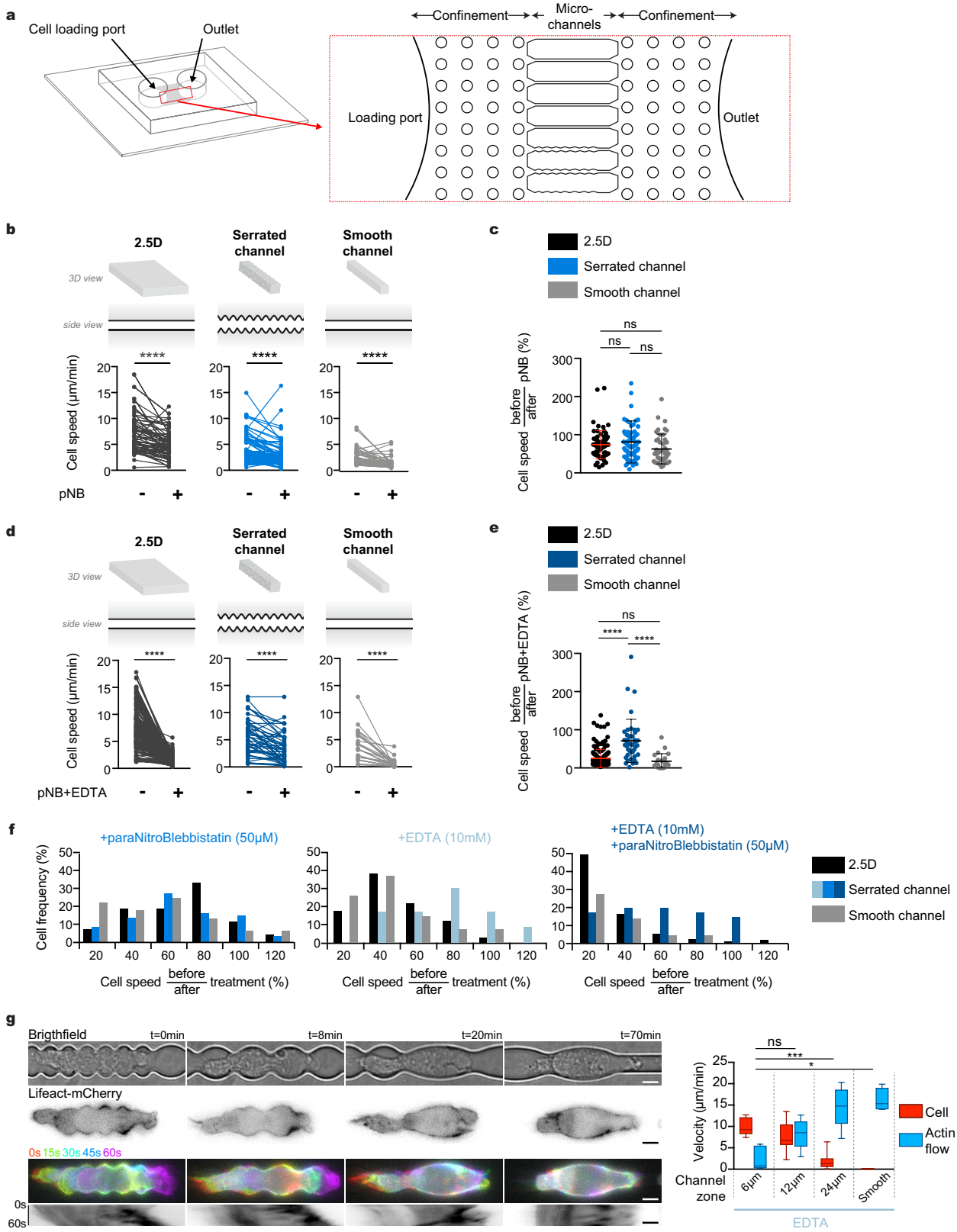
Extended Data Fig. 2 | Topography rescues T cell locomotion in the absence of adhesion. a–**c**, Migration of control ($n = 71$) and talin-KO ($n = 74$) cells under 3-µm confinement. Representative of four independent experiments. In **a**, the left panel shows the scheme of the 3-µm confinement used in **a**–**c**. The right panel shows representative snapshots of control and talin-KO cells under 3-µm confinement at $t = 60$ min. Individual tracks are displayed in different colours. Scale bars, 50 µm. In **b**, the speed and displacement of control and talin-KO cells under 3-µm confinement are shown. Data are mean ± s.d.; $^{****}P < 0.0001$, two-sided Mann–Whitney *U*-test. In **c**, the MSD of control cells (black) and talin-KO cells (red) is shown. Data are mean ± s.e.m.; $P < 0.0001$, two-sided Mann–Whitney *U*-test. **d**, Related to Fig. 2a–d. Scheme of multilayered microfluidic devices used to push cells into the pillar maze or into microchannels used in **e** and **f**. **e**, Snapshots of video microscopy at $t = 60$ min of control and talin-KO cells in the pillar device. The nucleus is shown in cyan (Hoechst), Lifeact-GFP reporter in red and the bright-field image in grey. Individual cell tracks are displayed in yellow. Scale bars, 50 µm. Time is in min:s. Representative of three experiments. **f**, Related to Fig. 2g, h. Velocities (µm/min) of control and talin-KO cells in serrated channels of different diameters (with a conserved serration period of 6 µm). Representative of three experiments. $n = 57$ for control cells and $n = 9$ for talin-KO cells. $^{*}P = 0.0356$; otherwise not significant, Kruskal–Wallis one-way ANOVA followed by post hoc Dunn's test. Boxes extend from the 25th to the 75th percentiles, with the middle line showing the median and the whiskers representing the minimum to maximum values.



Extended Data Fig. 3 | Topography rescues dendritic cell locomotion in the absence of adhesion. a, Top left, scheme of a mature dendritic cell migrating in a channel designed for the EDTA experiment. Mature dendritic cells migrate in a device containing channels with different serration periods, with or without 30 mM EDTA. Top right, single-cell migration speed with (+, blue) and without (-, grey) EDTA treatment. Bottom, histograms of cell speed in different serrations with (blue) or without (grey) EDTA. Representative of three independent experiments. $n = 35$ cells without EDTA, $n = 17$ cells with EDTA. Data are mean ± s.d.; $^{**}P = 0.0022$, $^{****}P < 0.0001$ and ns $P = 0.6363$, one-way ANOVA with Kruskal–Wallis test followed by post hoc Dunn's test. **b**, Top left, scheme of a mature dendritic cell migrating in a channel designed for the EDTA experiment, along a CCL19 gradient (indicated in red). Mature dendritic cells migrate towards the chemokine CCL19 into a device that contains microchannels with different serration periods, with or without 30 mM EDTA. Top right, migration speed with (+, blue) and without (-, grey) EDTA treatment. Bottom, histograms of cell speed in the different serrated channels with (blue) or without (grey) EDTA. Representative of three independent experiments. Without EDTA: $n = 44$ cells; with EDTA: $n = 52$ (6–12 μm), $n = 44$ (24 μm), $n = 33$ (smooth channels). Data are mean ± s.d.; $^{***}P = 0.0001$ and $^{****}P < 0.0001$, one-way ANOVA with Kruskal–Wallis test followed by post hoc Dunn's test.

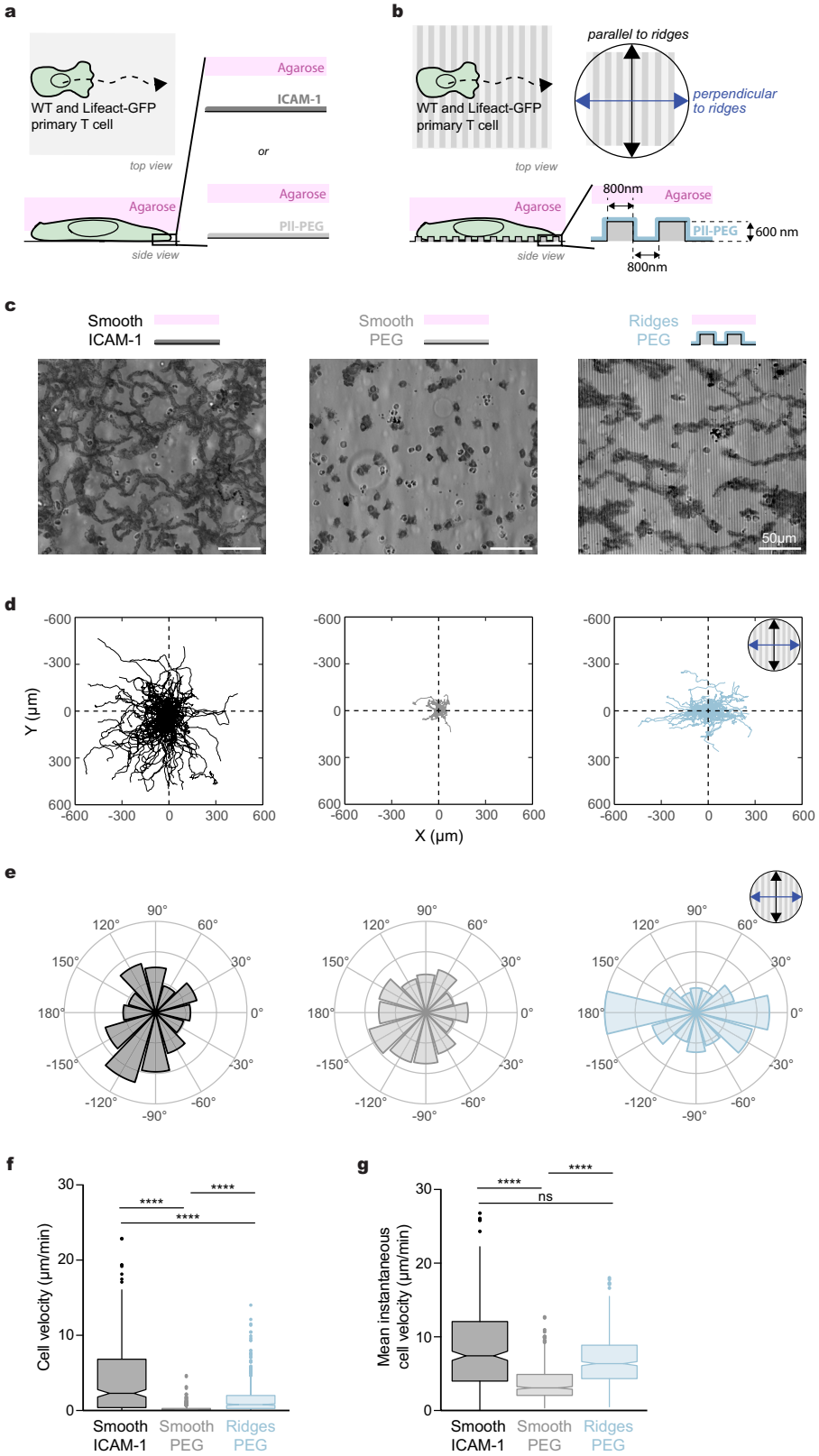


Extended Data Fig. 4 | Effect of topographical features on T cell migration. **a**, Related to Fig. 4d. Scheme of channel designs with decreasing complexity. In brief, serrations with a period of 6, 12 and 24 µm were arranged in consecutive zones of a 5-µm wide channel ending with a smooth zone. The left panel shows the different geometries. The right and middle panels show the velocity of cells migrating in zones with a different period. The total number of cells in each design is shown. Representative of four (control cells) and eight (talin-KO cells) independent experiments. Control: *P = 0.0415. KO arrow 'right': ***P = 0.0008 and **P = 0.0013; KO arrow 'left': ***P = 0.0003 and **P = 0.0018; KO bulb: *P = 0.0344; KO half-bulb: **P = 0.0016 (6 µm versus smooth) and **P = 0.0013 (12 µm versus smooth), otherwise not significant, one-way ANOVA with Kruskal–Wallis test followed by post hoc Dunn's test. Boxes extend from the 25th to the 75th percentiles, with the middle line showing the median and the whiskers representing the minimum to maximum values. **b**, Top, scheme of cell and actin retrograde flow length. Bottom, control cells expressing the Lifeact-GFP reporter with 10 mM EDTA (grey) or talin-KO cells (red), under 5-µm confinement (left, n = 12) or in a 5 × 5-µm channel (right, n = 9 for control cells and n = 8 for talin-KO cells), were observed by TIRF microscopy, and the mean cell length (ns P = 0.0555, two-sided Mann–Whitney U-test) and the actin retrograde flow length (ns P = 0.7247, two-sided Mann–Whitney U-test) were measured using kymograph analysis. Representative of two independent experiments. Data are mean ± s.d.



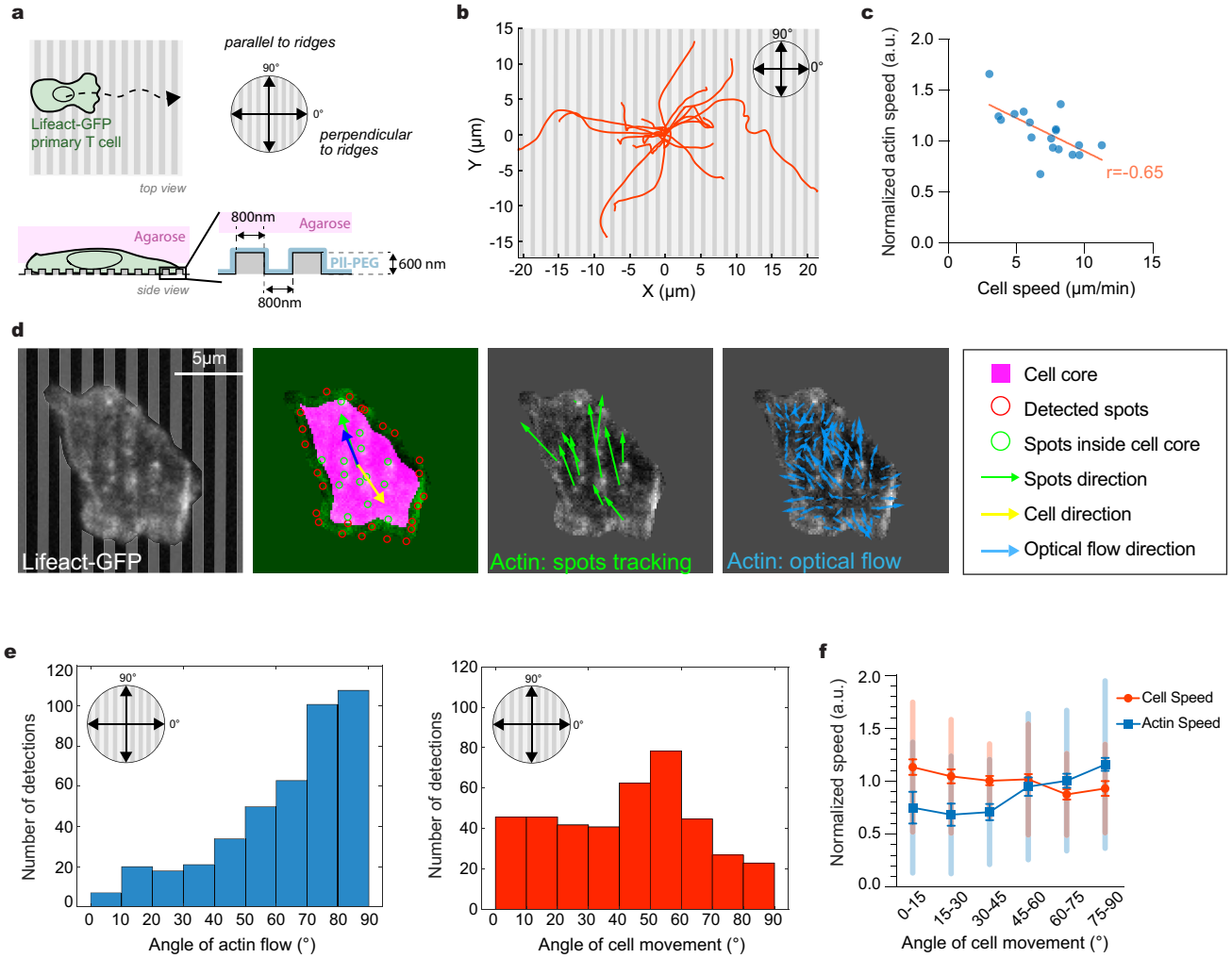
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Actin polymerization versus contractility during topography-based locomotion. a, Scheme of the device used for the experiments in Fig. 2i, j and Supplementary Figs. 3, 5. **b**, Control T cells migrating under confinement (2.5D) or in smooth versus serrated channels as shown in the top panels. After 60 min, 50 µM para-nitroblebbistatin (pNB) was added. The migration speed before (-) and after (+) treatment is shown. Representative of four independent experiments. $n = 68$ for 2.5D, $n = 79$ for serrated channel and $n = 44$ for smooth channel. ****$P < 0.0001$, two-tailed, Wilcoxon matched-pairs signed-rank test. **c**, Migration speed change after versus before pNB treatment. Data are mean ± s.d. ns $P > 0.9999$ (2.5D versus serrated channel), ns $P = 0.1414$ (2.5D versus smooth channel) and ns $P = 0.1044$ (serrated vs smooth channel), one-way ANOVA with Kruskal–Wallis test followed by post hoc Dunn's test. **d**, Control T cells migrating under confinement or in smooth versus serrated channels as shown in the top panels. After 60 min, 50 µM pNB and 10 mM EDTA were added. The migration speed before (-) and after (+) treatment is shown. Representative of four independent experiments. $n = 166$ for 2.5D, $n = 45$ for serrated channel and $n = 22$ for smooth channel. ****$P < 0.0001$, two-tailed, Wilcoxon matched-pairs signed-rank test. **e**, Migration speed change after versus before pNB + EDTA treatment. Data are mean ± s.d. ns $P = 0.1170$ and ****$P < 0.0001$, one-way ANOVA with Kruskal–Wallis test followed by post hoc Dunn's test. **f**, Histogram of migration speed variation after versus before pNB treatment alone (left, see **b**, **c**), EDTA treatment alone (middle, related to Fig. 2i, j) or pNB + EDTA treatment (right, see **d**, **e**). **g**, Related to Fig. 4e–g. Left, snapshots of a control T cell expressing Lifeact-mCherry incubated with 10 mM EDTA in different channel zones. The left panel from top to bottom shows bright field, Lifeact-mCherry in black, colour-coded snapshots of F-actin in the different channel zones (15-s intervals), and the corresponding kymograph of a cell migrating in the channel. Scale bars, 5 µm. Right, cell and actin retrograde flow velocities observed by wide-field microscopy and measured with kymograph analysis. $n = 8$ cells; 3 cells in all zones, 4 cells in the 6–12–24-µm zone and 1 cell in the 12–24-µm zone. *$P = 0.0138$ and ***$P = 0.0002$, one-way ANOVA with Kruskal–Wallis test followed by post hoc Dunn's test. Boxes extend from the 25th to the 75th percentiles, with the middle line showing the median and the whiskers representing the minimum to maximum values.

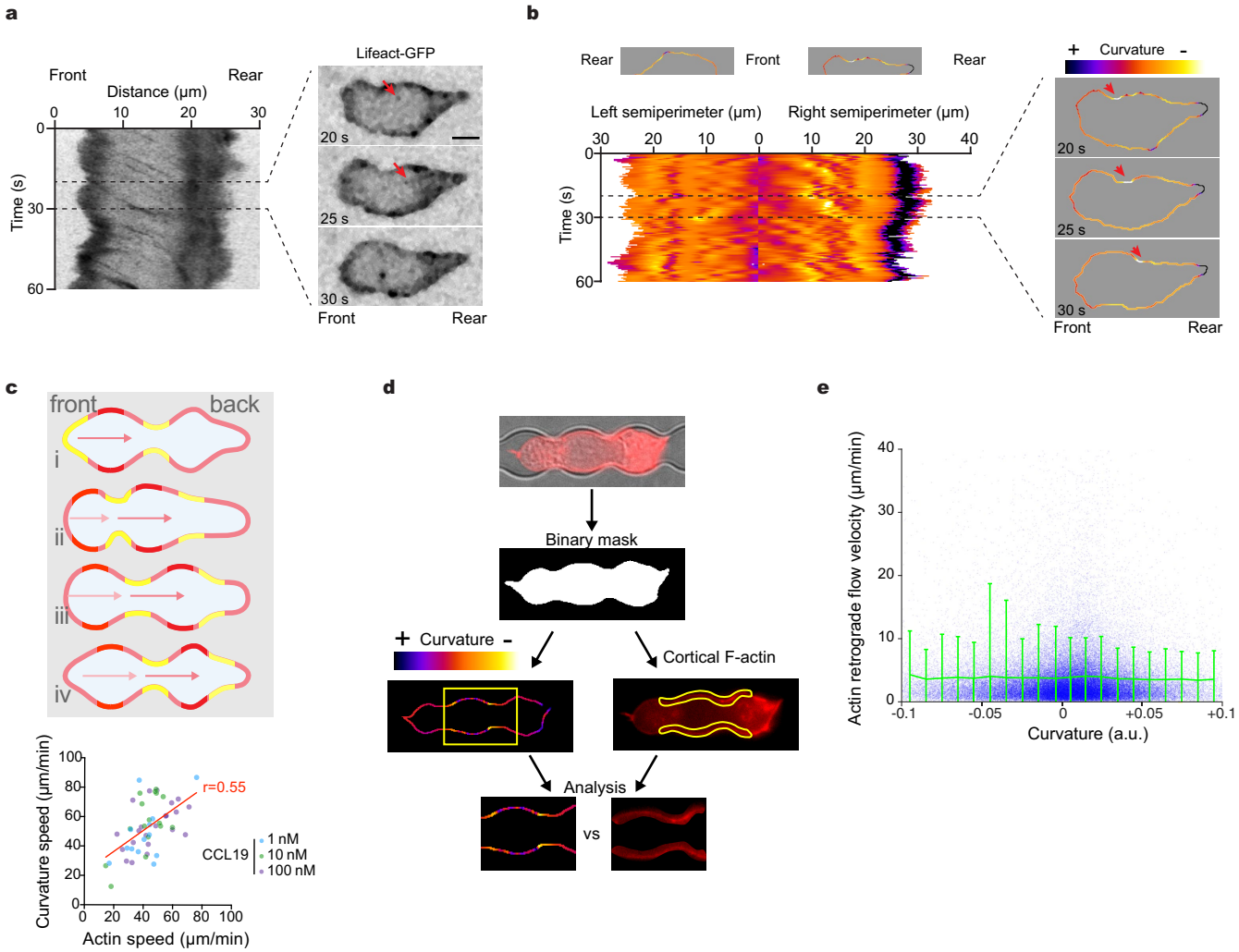


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Topography-based migration of primary T cells. **a**, Scheme (top and side view) of control and Lifeact-GFP cells confined under agarose on a smooth surface coated with Fc-ICAM1 or passivated with PLL-PEG. WT, wild type. **b**, Scheme (top and side view) of control and Lifeact-GFP cells confined under agarose on a coverslip with linear ridges passivated with PLL-PEG. **c**, Minimum-intensity projection showing the trace of control cells migrating for 15 min under agarose on Fc-ICAM1-coated smooth surfaces (left), PLL-PEG-coated smooth surfaces (middle) and PLL-PEG-coated ridge-baring surfaces (right). Representative of six independent experiments. Scale bars, 50 µm. **d**, Randomly selected tracks of mixed control and Lifeact-GFP cells migrating under agarose on Fc-ICAM1-coated smooth surfaces (left, $n = 183$), PLL-PEG-coated smooth surfaces (middle, $n = 183$) and PLL-PEG-coated ridge-baring surfaces (right, $n = 183$). The orientation of ridges is indicated in the top right corner. **e**, Rose diagram (with circular lines from the centre showing 6%, 12% and 18% of total cells) of cells migrating on Fc-ICAM1-coated smooth surfaces (left, $n = 419$), PLL-PEG-coated smooth surfaces (middle, $n = 650$) and PLL-PEG-coated ridge-baring surfaces (right, $n = 687$). The orientation of ridges is indicated in the top right corner. $n = 6$ independent experiments. **f**, **g**, Velocities of cells migrating under agarose on Fc-ICAM1-coated smooth surfaces ($n = 419$), PLL-PEG-coated smooth surfaces ($n = 650$) and PLL-PEG-coated ridge-baring surfaces ($n = 687$). $n = 6$ independent experiments. Cell velocities (cell displacement divided by track duration) (**f**) and mean instantaneous velocities (**g**) are shown. $^{****}P < 0.0001$ and ns $P = 0.0757$, both one-way ANOVA with Kruskal–Wallis test followed by post hoc Dunn's test.



Extended Data Fig. 7 | Actin-based force transmission in primary T cells. **a**, Scheme (top and side view) of cells migrating under agarose on a non-adhesive, topography-baring substrate. **b**, Tracks of migrating cells analysed in **c**–**f**. $n = 18$; mean cell speed = 6.58 µm/min; mean actin retrograde flow speed = 6.53 µm/min. Starting points of tracks are shifted to the origin. **c**, Mean speed of cells in relation to their mean actin retrograde flow speed (in relation to the substrate). Pearson's rank correlation coefficient r = −0.6484. A drop in retrograde flow with increasing locomotion speed demonstrates that, as in the transmembrane clutch paradigm of force transmission, actin slippage is inversely related to locomotion. **d**, Snapshots of actin flow analysis on ridged surfaces of cells expressing Lifeact-GFP. Left, cell core segmented with Ilastik (pink); the yellow arrow indicates cell direction, the blue arrow indicates the mean optical flow direction and the green arrow indicates actin retrograde flow. Middle, optical flow analysis (obtained with a customized MATLAB code). Right, actin single-spot tracking (TrackMate). Representative of three independent experiments. **e**, Angular distribution of actin retrograde flow and forwards locomotion of cells migrating on nanoridges. Only frames where optical flow analysis was confirming automatic particle tracking (deviation of <30°; see **d**) were subjected to analysis. The histogram on the left shows predominant retrograde actin flow when aligned along the ridges (90°). Retrograde actin flow declines towards 0° (perpendicular to the ridges), while forwards locomotion (right histogram) steadily increases, indicating that retrograde actin flow couples to topographical barriers and drives locomotion. $n = 422$ events of actin flow and cell movements obtained in $n = 18$ cells from three independent experiments. **f**, Frame-to-frame speed of cells migrating on nanoridges increases when forwards locomotion aligns perpendicular to topographical barriers (90° → 0° ≃ 15% increase). The orientation-dependent increase of cell speed is paralleled by a steady decrease of retrograde actin flow (90° → 0° ≃ 55%), indicating that retrograde actin flow couples to topographical barriers and drives locomotion. Frame-to-frame cell speed and actin retrograde flow speed were recorded in 18 cells from 3 independent experiments, and pooled to $n = 412$ (cell speed) and $n = 422$ (actin retrograde flow speed) events. The centre shows the mean, and both s.e.m. (solid lines) and s.d. (thick transparent lines) are shown. Before pooling, cell speed was normalized to the mean speed of the tracked cell, and actin retrograde flow speed was normalized to the mean actin retrograde flow speed.



Extended Data Fig. 8 | Retrograde actin flow and shape changes in adhesion-free confinement. Related to Fig. 4h. **a**–**c**, Primary T cells expressing Lifeact-GFP confined on PLL-PEG. $n = 51$ cells from three independent experiments: $n = 22$ cells for 100 nM CCL19, $n = 16$ cells for 10 nM CCL19 and $n = 13$ cells for 1 nM CCL19. **a**, Representative snapshots (right, $t = 20$, 25 and 30 s) and kymographs (left), of primary T cells expressing Lifeact-GFP confined on PLL-PEG, showing actin retrograde flow (red arrows). Scale bar, 5 µm. **b**, Representative snapshots (right, $t = 20$, 25 and 30 s) of the cell outline and the colour-coded curvature. Kymograph analysis (left) of the migrating T cell shows retrograde movement of cell body deformations (right, red arrowhead). **c**, Top, scheme of the travelling cell body deformations. Bottom, retrograde actin flow versus curvature flow velocities in primary T cells confined on an inert substrate and exposed to the indicated chemokine (CCL19) concentrations. $n = 51$ cells from three independent experiments: $n = 22$ cells for 100 nM CCL19, $n = 16$ cells for 10 nM CCL19 and $n = 13$ cells for 1 nM CCL19. Pearson's rank correlation coefficient $r = 0.5542$. **d**, **e**, Cell curvature versus cortical actin retrograde flow analysis in channels (obtained from three cells in three experiments). A scheme of the segmentation used to measure channel-dependent cell curvature and cortical actin retrograde flow velocity used in **e** is shown (**d**). Fluorescent images were binarized and the local curvature was calculated from a spline fit to the cell outline. Measurement of channel-dependent cell curvature versus cortical actin retrograde flow shows no correlation (Pearson's correlation coefficient $c \approx -0.006425$) (**e**). The green line shows mean ± s.d.

Corresponding author(s): Michael Sixt (michael.sixt@ist.ac.at)
Anne Reversat (anne.reversat@gmail.com)

Last updated by author(s): 2020/02/19

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collection for imaging was performed with NIS-Elements 5.02.00 (Nikon Eclipse microscope), VisiView 2.1.1 (Visitron TIRF/ epifluorescence microscope), Andor iQ 2.6 (Andor spinning-disc microscope), and iMIC Live Acquisition Software (TILL Photonics, FEI, Hillsboro, OR) ; see also the methods section 'Imaging'. Cell sorting was performed using FACS Diva 6.1.3. Design of microfluidic devices was performed with Coreldraw X8 (Corel Corporation, USA). A custom script was used to control the valves (Matlab, MathWorks, USA).
Data analysis	Analysis was performed with Fiji (ImageJ) 1.0; Matlab R2018a (MathWorks, USA); ilastik v1.1.5. Custom-made scripts for analysis (Matlab, MathWorks, USA) are available upon request. All statistical analyses were performed with GraphPad Prism 7 (GraphPad Software, USA).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data that support the findings of this study are available within the article and its Supplementary Information. Source Data for Figs. 1, 2, 4 and Extended Data Figs 1–8, are provided with the paper. Any additional information and related data are available from the corresponding authors upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We indicate in supplementary table Statistics and Reproducibility an exhaustive summary of all sample size and statistics for each experiment. No predeterminations of samples sizes were done; every experimental repetition included parallel imaging of multiple cells, resulting in samples sizes of mostly in the range of tens to hundreds.
Data exclusions	Data were only excluded in the case of microfluidic experiments based on one pre-established criterium: when microfluidic-based cell pushing was unsuccessful, such as valve leakage or high cell damage, the experiment was removed from analysis
Replication	Replications of all data was successful. All data were replicates at least 3 times, see supplementary table Statistics and Reproducibility for recapitulations of experiments.
Randomization	No randomization was performed : all cells assigned to control and experimental groups were analyzed.
Blinding	Investigators were not blinded to group allocation for majority of the study as data collection and analysis was performed by the same individual assigning the groups; however, key findings (e.g. cell velocities within different zones) were reproduced in an unbiased-(blinded) manner by the use of Fiji and custom-made analysis scripts.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Mouse Talin (anti-pan-talin, clone 8d4, T3287, Sigma-Aldrich, USA; dilution 1/200). Mouse GAPDH (clone GA1R, ab125247, Abcam, UK; dilution 1/3000). Goat anti-mouse IgG HRP conjugate (170-6516; Biorad, USA; dilution 1:5000)
Validation	All antibodies represent commercial standard validated antibody markers, validation data are available on vendor websites. https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Sigma/Datasheet/1/t3287dat.pdf https://www.abcam.com/gapdh-antibody-ga1r-loading-control-ab125247.pdf https://www.bio-rad.com/webroot/web/pdf/lsr/literature/LIT372.pdf

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	The LMR7.5 T cell hybridoma was a gift from A.M. Lennon-Duménil (Institut Curie, Paris, France) and created by N. Glaichenhaus (Université Côte d'Azur, Nice, France). The Lenti-X™ 293T Cell Line was obtained from TBUSA, formerly known as Clontech Laboratories, Inc. (632180). PLB-985 cells were obtained from the DSMZ (ACC 139)

Authentication	Cell line authentication was performed on the Lenti-X™ 293T by the seller. For the LMR7.5 T cell hybridoma, RNA sequencing was performed. PLB-985 (ACC-139) was validated by the DSMZ (cell surface markers, DNA fingerprint).
Mycoplasma contamination	All cell lines were tested and negative for mycoplasma.
Commonly misidentified lines (See ICLAC register)	No commonly mis-identifed cell line were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	6–12-week-old, male or female WT or Lifeact-GFP-expressing C57BL/6J mice were used for (1) primary T cell selection from lymph nodes and spleen, and (2) dendritic cell differentiation from bone marrow progenitors.
Wild animals	The study did not involved wild animals.
Field-collected samples	The study did not involve sample collection from the field.
Ethics oversight	All animal experiments at IST Austria were in accordance with Austrian law for animal experiments. Permission was granted by the Austrian Federal Ministry of Science, Research and Economy (identification code: BMWF-66.018/0005-II/3b/2012).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	All cell suspensions were filtered through a cotton mesh to remove cell aggregates and clumps, no staining was performed
Instrument	FACS Aria III was used for cell sorting (with nozzle 85µm and laser lines 488 and 561nm).
Software	Data were collected with FACS Diva 6.1.3.
Cell population abundance	Lifeact-GFP or Lifeact-mCherry positive cells were collected.
Gating strategy	Doublets and debris were excluded based on FSC and SSC characteristics.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Oncometabolites suppress DNA repair by disrupting local chromatin signalling

https://doi.org/10.1038/s41586-020-2363-0

Received: 1 February 2019

Accepted: 28 April 2020

Published online: 3 June 2020

Check for updates

Parker L. Sulkowski[1,2], Sebastian Oeck[1,3], Jonathan Dow[1,2], Nicholas G. Economos[1,2], Lily Mirfakhraie[4], Yanfeng Liu[1], Katelyn Noronha[1,5], Xun Bao[6], Jing Li[6], Brian M. Shuch[7], Megan C. King[4], Ranjit S. Bindra[1,8,9]✉ & Peter M. Glazer[1,2,9]✉

Deregulation of metabolism and disruption of genome integrity are hallmarks of cancer[1]. Increased levels of the metabolites 2-hydroxyglutarate, succinate and fumarate occur in human malignancies owing to somatic mutations in the isocitrate dehydrogenase-1 or -2 (*IDH1* or *IDH2*) genes, or germline mutations in the fumarate hydratase (*FH*) and succinate dehydrogenase genes (*SDHA*, *SDHB*, *SDHC* and *SDHD*), respectively[2–4]. Recent work has made an unexpected connection between these metabolites and DNA repair by showing that they suppress the pathway of homology-dependent repair (HDR)[5,6] and confer an exquisite sensitivity to inhibitors of poly (ADP-ribose) polymerase (PARP) that are being tested in clinical trials. However, the mechanism by which these oncometabolites inhibit HDR remains poorly understood. Here we determine the pathway by which these metabolites disrupt DNA repair. We show that oncometabolite-induced inhibition of the lysine demethylase KDM4B results in aberrant hypermethylation of histone 3 lysine 9 (H3K9) at loci surrounding DNA breaks, masking a local H3K9 trimethylation signal that is essential for the proper execution of HDR. Consequently, recruitment of TIP60 and ATM, two key proximal HDR factors, is substantially impaired at DNA breaks, with reduced end resection and diminished recruitment of downstream repair factors. These findings provide a mechanistic basis for oncometabolite-induced HDR suppression and may guide effective strategies to exploit these defects for therapeutic gain.

The oncometabolites 2-hydroxyglutarate (2HG), succinate and fumarate inhibit α-ketoglutarate (αKG)-dependent dioxygenases[7,8] including histone lysine demethylases and other epigenetic modifiers[9–11]. We initially identified two lysine demethylases, KDM4A and KDM4B, as potential targets for oncometabolite suppression of HDR in an initial screen[5]. To further investigate this, we assembled a series of human cancer cell lines with endogenous and engineered mutations in IDH1 or IDH2, FH and SDH, short hairpin RNA (shRNA) knockdowns, and CRISPR modifications and confirmed the expected levels of oncometabolites and the corresponding hypermethylation of H3K9—a target for demethylation by KDM4A and KDM4B (Extended Data Fig. 1a–h).

On the basis of the capacity of KDM4A or KDM4B to regulate gene expression[11,12], we considered that 2HG, fumarate and succinate might suppress HDR via gene downregulation. Transcriptome analyses comparing *IDH1*$^{R132H/+}$ cells to wild-type (*IDH1*$^{+/+}$) cells showed broad changes in gene expression (Extended Data Fig. 1i), but no correlation of IDH1 status with HDR genes (Extended Data Fig. 1j). Gene expression patterns in human gliomas in The Cancer Genome Atlas (TCGA) lower grade glioma cohort indicated that HDR genes are not suppressed in mutant IDH tumours (Extended Data Fig. 1k). Western blot analyses of isogenic cell lines with or without R132H mutant IDH1 or *FH* or *SDH* shRNA knockdown showed no differences in the HDR factors RAD51, BRCA2, ATM, TIP60, MRE11 or RPA (Extended Data Fig. 1l).

Alternatively, we tested for a direct functional effect on HDR. We observed increased levels of double-strand breaks (DSBs), which are characteristic of HDR-deficient cells[5,6], as early as 2 h after the addition of 2HG, fumarate or succinate to cells (Fig. 1a, Extended Data Fig. 1m, n). We confirmed that intracellular levels of the metabolites and H3K9me3 were increased at the 2-h time point, indicating inhibition of KDM4A or KDM4B (Extended Data Fig. 1o–t). Such rapid kinetics pointed to a direct effect of the metabolites on HDR rather than on gene expression.

We next examined the formation of DNA repair foci by RAD51 and BRCA1 in response to ionizing radiation. In control cells, RAD51 foci are detectable at 2 h after ionizing radiation, peaking at 4–6 h (Fig. 1b, c, Extended Data Fig. 2a,b), consistent with previous studies[13]. But RAD51 foci were attenuated in cells treated with fumarate or succinate and in cells with shFH or shSDH knockdown (Fig. 1b, c, Extended Data Fig. 2c–e). We also compared RAD51 foci in cells with or without mutant IDH1 or IDH2 (including CRISPR–Cas9-mediated knockout of the endogenous *IDH1*^{R132C} allele in HT1080 fibrosarcoma cells), with or without 2HG

[1]Department of Therapeutic Radiology, Yale University School of Medicine, New Haven, CT, USA. [2]Department of Genetics, Yale University School of Medicine, New Haven, CT, USA. [3]Department of Medical Oncology, University of Duisburg-Essen, Essen, Germany. [4]Department of Cell Biology, Yale University School of Medicine, New Haven, CT, USA. [5]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. [6]Karmanos Cancer Institute, Wayne State University, Detroit, MI, USA. [7]Department of Urology, University of California at Los Angeles, Los Angeles, CA, USA. [8]Department of Pathology, Yale University School of Medicine, New Haven, CT, USA. [9]These authors jointly supervised this work: Ranjit S. Bindra, Peter M. Glazer. ✉e-mail: ranjit.bindra@yale.edu; peter.glazer@yale.edu

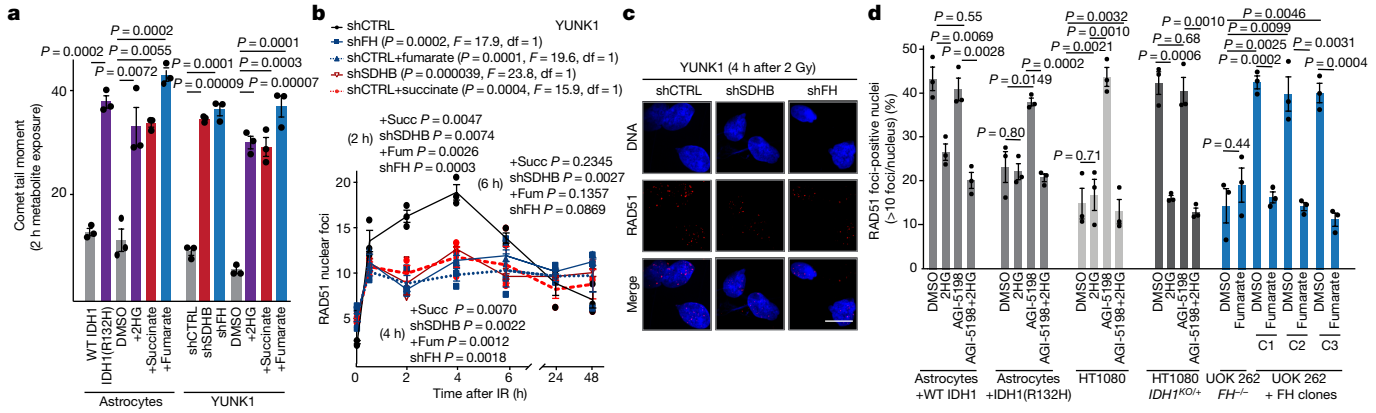


Fig. 1 | Oncometabolites directly suppress HDR. a, Quantification of neutral comet assays performed in immortalized astrocytes overexpressing wild-type (WT) IDH1 or IDH1(R132H) or treated with 2HG, succinate or fumarate, and in YUNK1 cells after shRNA suppression of *FH* or *SDHB* (shFH or shSDHB, respectively) or the addition of 2HG, succinate or fumarate. shCTRL, non-targeting control shRNA. Dimethylsulfoxide (DMSO) was used a vehicle control. **b**, Quantification (**b**) and representative images (**c**) of RAD51 nuclear foci at the indicated time points after 2 Gy ionizing radiation (IR) treatment in YUNK1 cells after shRNA suppression of *FH* or *SDHB*, or after pre-treatment with fumarate (Fum) or succinate (Succ). Scale bar, 10 µm. **d**, Quantification of cells with RAD51 foci-positive nuclei in immortalized astrocytes overexpressing wild-type IDH1 or IDH1(R132H), and in HT1080 fibrosarcoma cells ($IDH1^{R132C/+}$), and in HT1080 cells with CRISPR–Cas9-mediated knockout of the $IDH1^{R132C}$ allele ($IDH1^{KO/+}$). Cells were treated with or without 2HG for 24 h, 1 µM AGI-5198 for 5 days or a combination thereof, before irradiation. Data are mean ± s.e.m. from $n = 3$ biological replicates. *P* values determined by two-tailed unpaired *t*-test, df = 4 (**a**, **d**), or analysis of variance (ANOVA) (**b**).

treatment, and with or without treatment with the mutant IDH1- or IDH2-specific inhibitors AGI-5198 or AG-221, respectively. The formation of RAD51 foci was consistently impaired by increased 2HG (Fig. 1d, Extended Data Fig. 2f, g). $FH^{-/-}$ UOK 262 human renal cell carcinoma cells showed low levels of RAD51 foci formation, but this was rescued by FH complementation, and then suppressed again when fumarate was added back (Fig. 1d). The formation of BRCA1 foci was similarly impaired by high levels of metabolites (Extended Data Fig. 2h–m). Consistent with HDR impairment, mutant IDH1 cells and tumours were radiosensitive compared to isogenic controls (Extended Data Fig. 2n–p).

Next, we developed a ligand-dependent double-strand break and chromatin immunoprecipitation (DSB–ChIP) assay (Fig. 2a) to monitor recruitment kinetics of HDR factors to a site-specific DSB. In this assay, a site-specific DSB is induced in U2OS DR-GFP cells[14,15] (a human osteosarcoma cell line with a chromosomally integrated green fluorescent protein (GFP)-based HDR reporter) by a ligand-responsive I-SceI nuclease. The timing of DSB induction is precisely controlled in these cells, because the addition of the ligands shield-1 and triamcinolone rapidly causes I-SceI stabilization and nuclear localization, inducing the site-specific DSB within the DR-GFP reporter gene locus[15] (Fig. 2a, Extended Data Fig. 3a).

Cells in the presence or absence of the respective metabolites were treated with shield-1 and triamcinolone to rapidly induce the site-specific DSBs, and ChIP analyses were performed over time. Figure 2b presents heat maps with the ChIP results normalized to uninduced cells at time zero for each respective antibody (with additional quantification in Extended Data Fig. 3b–l and antibody validation in Extended Data Fig. 3m–u). Phosphorylated γH2AX, a marker of DSBs, rapidly accumulated at the DSB in both control and metabolite-treated cells. However, the γH2AX signal resolved quickly in control cells but persisted in metabolite-treated cells, suggesting reduced repair. In control cells, we detected a rapid spike in H3K9 trimethylation at the DSB, occurring within 30 min after DSB induction and disappearing after 1 h. This H3K9me3 spike was concurrent with the recruitment of the known H3K9 histone methyltransferase, SUV39H1. The H3K9me3 spike was followed by a coordinated pattern of DSB repair factor recruitment, beginning with MRE11, TIP60 and ATM, followed by RPA (indicating end-resection) and finally by BRCA1 and RAD51. By contrast, in cells with high levels of 2HG, fumarate or succinate, H3K9me3 at the site was already increased at time zero, before DSB induction, and remained increased, whereas the subsequent recruitment of the HDR factors was substantially attenuated. Notably, although H3K9me3 levels show a rapid spike at the induced DSB in control cells, in oncometabolite-treated cells the pre-existing and persistent H3K9 hypermethylation masks any potential spike (further quantified in Fig. 2c).

Cell cycle profiling showed no G1 stalling in cells with increased oncometabolites (Extended Data Fig. 4a–c), whereas growth rates were slightly slower than in controls (Extended Data Fig. 4d). On the basis of foci formation and functional assays, neither non-homologous end-joining (NHEJ) nor the micro-homology-mediated end joining (MMEJ) pathways were suppressed by increased metabolites (Extended Data Fig. 4e–n).

HDR is dependent on end-resection of DNA at the DSB to generate single-stranded DNA (ssDNA). To examine end-resection, we used a site-specific, quantitative PCR (qPCR) method[16] to measure ssDNA production after activation of I-SceI (Fig. 2d). In control cells, we found that at 3 h after DSB induction, there is protection of a qPCR product approximately 450 base pairs from the I-SceI site, consistent with end-resection (Fig. 2e). However, treatment of cells with 2HG, succinate or fumarate reduced this protection, which suggests a defect in end-resection.

We next used laser micro-irradiation to monitor histone dynamics at DSBs induced in a stripe across the nucleus. In cells with increased levels of 2HG, we observed high background levels of H3K9me3 in unirradiated nuclei versus low levels in wild-type cells (Fig. 2f, Extended Data Fig. 5a). After laser micro-irradiation, we observed that H3K9me3 was locally deposited at the stripe in wild-type cells, which was not seen in cells with increased levels 2HG. Similar results were seen in other cell lines with or without increased 2HG, fumarate or succinate (Extended Data Fig. 5b–m).

Activation of TIP60 at DSBs is known to be dependent on local H3K9 trimethylation, leading to TIP60-mediated activation of ATM[17–19]. We further evaluated the H3K9me3–TIP60–ATM recognition and signalling axis[18,20] and found that cells with increased metabolites (mutant IDH1, $FH^{-/-}$, shFH and shSDHB and/or treatment with metabolites) have a marked defect in TIP60 and phosphorylated-ATM (pATM) foci formation after ionizing radiation (Fig. 2g, Extended Data Fig. 6a–h).

Cells with increased metabolites also showed deficient ATM activation based on reduced phosphorylation of ATM at Ser1981 after

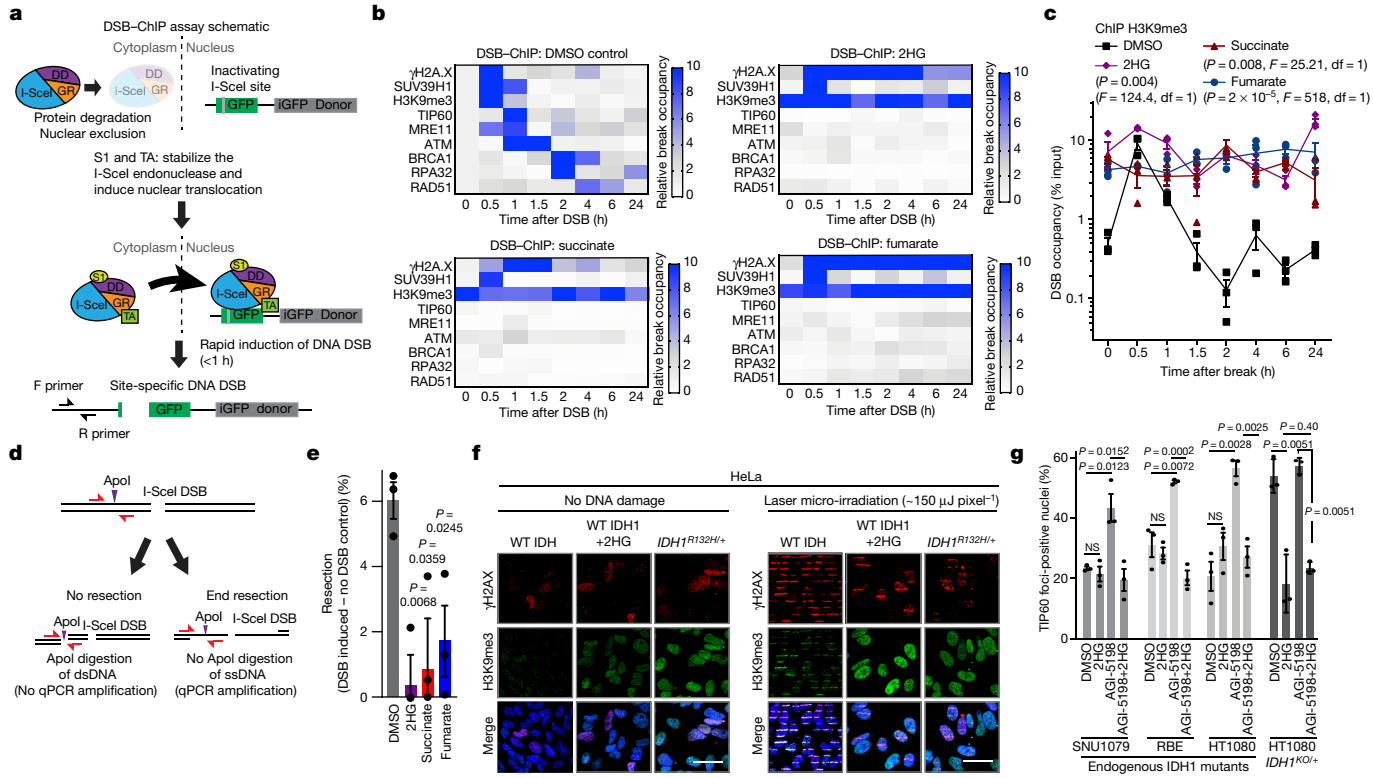


Fig. 2 | Oncometabolites suppress the stepwise recruitment of HDR factors to DNA DSBs. a, Schematic representation of the DSB–ChIP assay system. S1, shield-1; TA, triamcinolone. **b**, Heat maps of the relative occupancy of the indicated factors at the site-directed DSB, as measured by ChIP, and normalized to the uninduced controls at the indicated time points after the addition of shield-1 and triamcinolone. **c**, DSB–ChIP quantification over time of H3K9me3 levels at the site-specific DSB. **d**, Schematic representation of the PCR-based DNA end-resection assay. **e**, Quantification of end-resection at the same site-specific DSB as used in the DSB–ChIP assay. **f**, Immunofluorescence images of γH2AX and H3K9me3 in cell nuclei in the indicated cells without DNA damage or 1 min after laser micro-irradiation to induce DNA damage. Scale bars, 20 μm. **g**, Quantification of TIP60 foci-positive nuclei after ionizing radiation in HT1080 fibrosarcoma cells, and in HT1080 cells with a CRISPR–Cas9-mediated knockout of the *IDH1*^{R132C} allele as well as in SNU1079 (*IDH1*$^{R132C/+}$) and RBE (*IDH1*$^{R132S/+}$) cholangiocarcinoma cells treated as indicated with or without 2HG for 24 h, 1 μM AGI-5198 for 5 days or a combination thereof, before ionizing radiation. Data are mean ± s.e.m. from three biological replicates. *P* values determined by ANOVA (**c**) or two-tailed unpaired *t*-test, df = 4 (**e**, **g**).

ionizing radiation (Extended Data Fig. 6i). We also found a defect in phosphorylation at S*Q motif sites (targets for ATM) (Extended Data Fig. 6j). pATM foci were highly colocalized with TIP60 foci (80%) in control cells (Extended Data Fig. 6k), but not in cells treated with 2HG, fumarate or succinate (Extended Data Fig. 6k). Levels of phospho-RPA (pRPA), phospho-ATR (pATR) and phospho-CHK1 (pCHK1) were also reduced in oncometabolite producing cells (Extended Data Fig. 6l). CHK2 phosphorylation was partially reduced.

Next, we tested the effect of exogenous αKG to out-compete the increased metabolites and restore the function of αKG-dependent dioxygenase[7]. We found that αKG treatment of oncometabolite-producing cells decreased global H3K9 trimethylation (Extended Data Fig. 7a, b) and restored TIP60, pATM and RAD51 foci formation (Extended Data Fig. 7c–h). Treatment with αKG suppressed H3K9me3 hypermethylation by increased 2HG, succinate and fumarate on western blot (Extended Data Fig. 7i), and on ChIP analysis at the I-SceI locus in the absence of a break (Extended Data Fig. 7j). Treatment with αKG also restored the spike in H3K9me3 and recruitment of HDR factors at the DNA DSB (Extended Data Fig. 7k–aa), resulting in functional restoration of HDR in the DR-GFP assay (Extended Data Fig. 7ab).

In other rescue experiments in endogenous mutant IDH1 and *FH*$^{-/-}$ cells, overexpression of KDMA and KDM4B, but not the catalytically inactive mutants, KDM4A(H188A)[21] and KDM4B(H189A)[22], suppressed H3K9 hypermethylation, restored TIP60 and RAD51 foci, and suppressed the increased DSBs in the comet assay (Fig. 3a, Extended Data Fig. 8a–j). Overexpression of KDM4A or KDM4B also suppressed H3K9 hypermethylation (Extended Data Fig. 8k), reduced the comet tails (Extended Data Fig. 8l), and restored BRCA1 and RAD51 foci (Extended Data Fig. 8m–p) in other sets of oncometabolite-producing cells.

By contrast, there was no detectable rescue of any of these endpoints by overexpression of other αKG-dependent dioxygenases[7,8,23], including KDM4C, KDM6A, ALKBH2, ALKBH3 and JMJD4 (Fig. 3a, Extended Data Fig. 8a-j). We also ruled out a role for HIF-1 (Extended Data Fig. 8q, r).

We next generated individual KDM4A and KDM4B knockout cells in the YUNK1 cell background (Extended Data Fig. 8s). Notably, knockout of KMD4B, but not of KDM4A, induced high levels of H3K9 trimethylation (Extended Data Fig. 8s), impaired TIP60 and RAD51 foci formation (Fig. 3b, Extended Data Fig. 8t) and caused persistence of DSBs (Extended Data Fig. 8u). Forced expression of KDM4A or KDM4B rescued all these endpoints in KDM4B-knockout cells (Fig. 3b, Extended Data Fig. 8t–v).

Next, we conducted experiments knocking down KDM4A and KDM4B in either the KDM4A- or KDM4B-knockout cells as well as the parental control cells (Fig. 3c). We found that only KDM4B loss, either by knockdown or knockout, induces an HDR deficiency as measured by TIP60 and RAD51 foci formation or the comet assay (Fig. 3d, Extended Data Fig. 9a, b). To link the effect of KDM4B loss on HDR to its catalytic activity, we complemented KDM4B-knockout cells with either wild-type *KDM4B* cDNA or the catalytically inactive KDM4B(H189A) variant[22] (Fig. 3e). Wild-type *KDM4B*, but not KDM4B(H189A), suppressed H3K9me3 hypermethylation (Fig. 3e), rescued TIP60 and RAD51 foci (Fig. 3f, Extended Data Fig. 9c) and reduced DSBs (Extended Data Fig. 9d), indicating the catalytic activity of KDM4B is necessary for HDR.

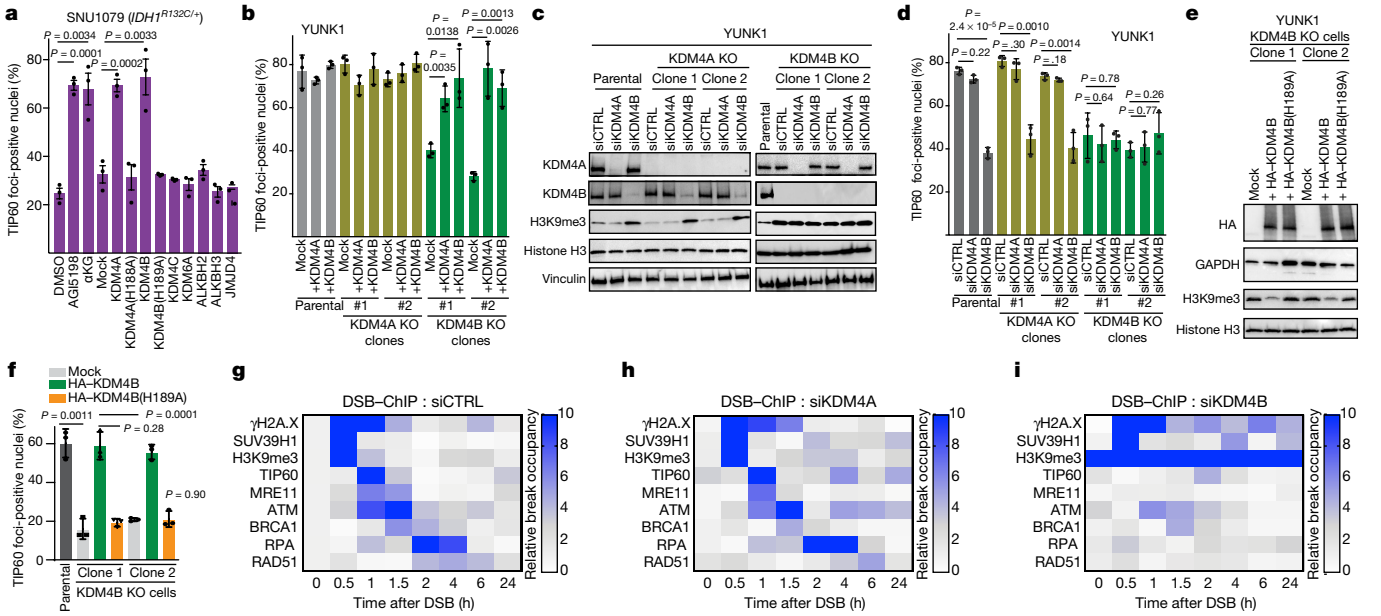


Fig. 3 | Oncometabolites suppress HDR via inhibition of KDM4B. a, Quantification of TIP60 foci-positive nuclei in SNU1079 (*IDH1*$^{R132C/+}$) cells exposed to ionizing radiation after transfection with expression vectors for the indicated αKG-dependent dioxygenases compared to treatment with 2 mM αKG or 1 μM AGI-5198. **c**, **d**, Western blot analysis (**c**) and quantification of TIP60 foci (1 h after 2 Gy ionizing radiation) (**d**) in *KDM4A*- or *KDM4B*-knockout YUNK1 cells after transfection with the indicated siRNAs. **e**, **f**, Western blot analysis (**e**) and quantification of TIP60 foci (**f**) after ionizing radiation in YUNK1 cells after transfection with expression constructs for wild-type KDM4B or catalytically inactive KDM4B(H189A). **g**–**i**, Heat map representation of DSB–ChIP assays in U2OS cells after transfection with non-targeting control siRNA (**g**), siKDM4A (**h**) or siKDM4B (**i**). Data are mean ± s.e.m. from three biological replicates. *P* values were determined by two-tailed unpaired *t*-test, df = 4. Western blots in **c** and **e** were performed twice with similar results.

Consistent with the data from the YUNK1 cells, knockdown of KDM4B, but not KDM4A, increased the overall level of H3K9me3 across the genome and at the reporter locus in the U2OS DSB ChIP cells (Extended Data Fig. 9e, f), impaired the recruitment of repair factors to the DSB in the DSB–ChIP assay (Fig. 3g–i, Extended Data Fig. 9g–o), and produced an increase in comet tails (Extended Data Fig. 9p), similar to the effects of the oncometabolites. These results connect loss of KDM4B activity to HDR deficiency and show that although forced overexpression of KD4MA can compensate for KDM4B loss, physiological levels of KDM4A cannot compensate for KDM4B loss.

To directly test KDM4B loss of function as the key mediator in oncometabolite-induced HDR deficiency, we conducted a series of epistasis experiments. We found that loss of KDM4B, but not KDM4A, either by knockout or short interfering RNA (siRNA) suppression, reduced HDR and was epistatic to 2HG produced by mutant IDH1, specifically placing KDMB in the pathway of oncometabolite-mediated HDR suppression (Extended Data Fig. 9q–w). Consistent with a KDM4B-associated HDR defect, we detected a decrease in cell proliferation rate with KDM4B knockout (Extended Data Fig. 9x–z). Notably, KDM4B-knockout mice are born at a normal Mendelian ratio[24], so although KDM4B inhibition affects HDR in human cells, it is not an essential gene in mice, similar to results seen with ATM knockout[25].

To experimentally uncouple histone hypermethylation at H3K9 from KDM4B inhibition, we expressed the histone mutant H3K9M, which acts in *trans* to reduce genomic H3K9me3 formation by sequestering the histone methyltransferases that can act on H3K9[26] (Extended Data Fig. 9aa). We also compared H3K9G, H3K9R and H3K9R mutants (Fig. 4a) along with similar constructs for other H3 lysine residues (H3K4, H3K27 and H3K36), because increased metabolites also cause hypermethylation of these lysine residues[7,8,10] (Extended Data Fig. 9ab). Only H3K9M expression reduced the H3K9me3 levels in oncometabolite-producing cells (Extended Data Fig. 9aa) and did so without affecting oncometabolite levels (Extended Data Fig. 9ac), meaning that KDM4B is still inhibited[7].

We found that only expression of H3K9M could restore TIP60 foci (Fig. 4a) and suppress comet tails in mIDH1 and FH-deficient cells (Extended Data Fig. 10a, b). H3K9M expression also suppressed comet tails, restored H3K9me3 levels, and increased ATM activation in SDHB- and FH-deficient cells (Extended Data Fig. 10c–e). Notably, we did not observe a rapid DSB-dependent spike in trimethylation at any of these other H3 lysine residues as we do with H3K9me3 (Extended Data Fig. 10f).

H3K9M expression in the U2OS DSB–ChIP cells suppressed H3K9me3 hypermethylation in spite of the presence of high levels of 2HG, both globally (Extended Data Fig. 10g) and at the DSB locus (Extended Data Fig. 10h). H3K9M expression also restored the local H3K9me3 methylation spike at the DSB and the recruitment of HDR factors, in spite of high 2HG (Fig. 4b–e, Extended Data Fig. 10i–q). These results support the hypothesis that increased levels of oncometabolites inhibit HDR because they increase baseline H3K9me3 hypermethylation, masking the ability of the cell to generate an H3K9me3 spike at a DSB, causing impaired recruitment of repair factors.

With H3K9M expression in cells with increased 2HG (Fig. 4e), the H3K9me3 ChIP signal disappears from the break over time (as it does in control cells) even though KDM4B demethylase activity is inhibited by the 2HG. To explain this, we note that this loss of the H3K9me3 signal correlates with loss of total H3 from the locus (Extended Data Fig. 10r), consistent with histone eviction occurring with end-resection during repair. Notably, we found that KDM4A and KDM4B are recruited to the DSB (Extended Data Fig. 10s). Because the experiments above show that local demethylation is not required once HDR is initiated, this may reflect the established role for these factors in competition with 53BP1 for binding to H4K20me2 at the break[27].

We also tested for the effect of H3K9R on HDR in wild-type cells, because it mimics an unmodifiable lysine residue. We found that H3K9R expression in wild-type cells produces increased comet tails and suppression of TIP60 and RAD51 foci (Extended Data Fig. 10t–v), indicating an induced HDR defect. H3K9R also sensitized wild-type U87 cells to

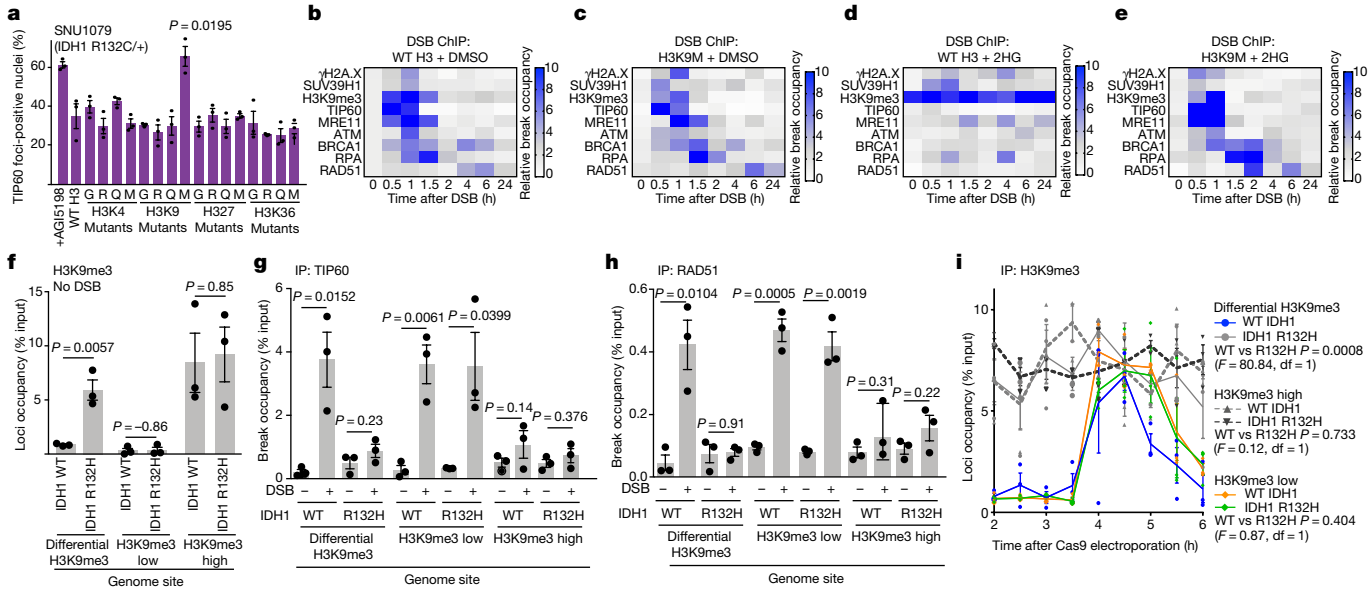


Fig. 4 | Aberrant H3K9 methylation impairs HDR and mechanistically underlies oncometabolite induced HDR deficiency. a, Quantification of TIP60 foci-positive nuclei in SNU1079 cells after expression of the indicated H3 mutant constructs or after treatment with the mIDH1 inhibitor, AGI-5198. **b**–**e**, Heat map representation of DSB–ChIP assays after transfection with either wild-type H3 or H3K9M and treatment with either DMSO control or 2HG, as indicated. **f**, Confirmation by ChIP of H3K9me3 levels in wild-type and IDH1(R132H)-expressing astrocytes at selected loci identified by analysis of published ChIP–seq data to show either differential H3K9me3 levels (high H3K9me3 in IDH1(R132H)-expressing astrocytes and low in wild-type astrocytes), low H3K9me3 levels in both cell lines, or high H3K9me3 levels in both. **g**, **h**, ChIP analysis of TIP60 (**g**) and RAD51 (**h**) recruitment 12 h after Cas9–guide RNA nucleofection to generate site-specific DSBs at the loci profiled in **f** in the wild-type IDH1 and IDH1(R132H) astrocyte cell lines. **i**, ChIP analysis over time of H3K9me3 levels at the same loci as in **f**–**h** after Cas9–guide RNA nucleofection. Data are mean ± s.e.m. from three biological replicates. *P* values were determined by two-tailed unpaired *t*-test, df = 4 (**a**, **f**–**h**) or ANOVA (**i**).

PARP inhibition (Extended Data Fig. 10w), but it did not have any further effect on the inherent PARP inhibitor sensitivity in mutant IDH1 cells (demonstrating epistasis). By contrast, H3K9M expression rescued the PARP inhibitor sensitivity in mIDH1 cells.

We next tested whether HDR might still be able to occur at sites in which the H3K9me3 levels were low at baseline even in cells with high 2HG. By analysis of published ChIP followed by high-throughput sequencing (ChIP–seq) data generated from the IDH1(R132H) mutant and wild-type astrocytes[28] (Extended Data Fig. 10x), we identified three types of genomic sites in the mutant IDH1 and wild-type cells: (1) hypomethylated at baseline in wild-type cells but hypermethylated in mutant IDH1 cells; (2) hypomethylated in both; and (3) hypermethylated in both, confirmed by ChIP (Fig. 4f). Using CRISPR–Cas9 to induce site-specific DSBs at these sites (Extended Data Fig. 10y), we assayed for the recruitment of DNA repair factors via ChIP. In the differentially H3K9 methylated site, we saw recruitment of HDR factors by ChIP in the wild-type but not mutant IDH1 astrocytes (Fig. 4g, h, Extended Data Fig. 10z, aa). At the site with low H3K9me3 in both, we observed recruitment of HDR factors in both cases. Notably, at this site, an H3K9me3 signal could be produced at the DSB not only in the wild-type astrocytes, but also in the IDH1 mutant ones (Fig. 4i). At the site with high H3K9me3 in both, there was no H3K9me3 signal above background in either the wild-type or mutant IDH1 cells and recruitment of HDR factors was attenuated. These results support our model that it is the pre-existing H3K9me3 hypermethylation (caused by metabolite inhibition of KDM4B) that inhibits repair of DSBs by the HDR pathway because the H3K9me3 signal cannot be generated. Consistent with the need for an H3K9me3 spike, siRNA knockdown of the H3K9 histone methyltransferase, SUV39H1, impaired HDR (Extended Data Fig. 10ab–ad), consistent with a previous report[17].

The above experiments provide a direct link between oncometabolite-induced HDR deficiency, KDM4B inhibition and H3K9me3 status. They identify H3K9 methylation as the key regulatory target in the pathway by which increased metabolites cause decreased HDR. This pathway is characterized in normal cells by a rapid spike in H3K9 trimethylation that coordinates recruitment of TIP60 and MRE11, promoting ATM activation, licensing end-resection, and leading to the downstream recruitment of RPA, BRCA1 and RAD51. In cells that overproduce metabolites, much of the genome has high levels of H3K9me3, and at such regions, this H3K9me3 signal is not properly induced. The constitutively high levels of H3K9me3 prevent a hypermethylation spike, which impairs HDR factor recruitment and end-resection, conferring sensitivity to PARP inhibitors. These results reveal a pathway by which metabolism influences DNA repair and may provide the basis for therapeutic strategies for patients with metabolite-associated malignancies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2363-0.

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
2. Dang, L. et al. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* **462**, 739–744 (2009).
3. Toro, J. R. et al. Mutations in the *fumarate hydratase* gene cause hereditary leiomyomatosis and renal cell cancer in families in North America. *Am. J. Hum. Genet.* **73**, 95–106 (2003).
4. Pollard, P. J. et al. Accumulation of Krebs cycle intermediates and over-expression of HIF1α in tumours which result from germline *FH* and *SDH* mutations. *Hum. Mol. Genet.* **14**, 2231–2239 (2005).
5. Sulkowski, P. L. et al. 2-Hydroxyglutarate produced by neomorphic IDH mutations suppresses homologous recombination and induces PARP inhibitor sensitivity. *Sci. Transl. Med.* **9**, eaal2463 (2017).
6. Sulkowski, P. L. et al. Krebs-cycle-deficient hereditary cancer syndromes are defined by defects in homologous-recombination DNA repair. *Nat. Genet.* **50**, 1086–1092 (2018).

7. Xu, W. et al. Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. *Cancer Cell* **19**, 17–30 (2011).
8. Xiao, M. et al. Inhibition of α-KG-dependent histone and DNA demethylases by fumarate and succinate that are accumulated in mutations of FH and SDH tumor suppressors. *Genes Dev*. **26**, 1326–1338 (2012).
9. Turcan, S. et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483**, 479–483 (2012).
10. Lu, C. et al. IDH mutation impairs histone demethylation and results in a block to cell differentiation. *Nature* **483**, 474–478 (2012).
11. Losman, J. A. & Kaelin, W. G. Jr. What a difference a hydroxyl makes: mutant IDH, (*R*)-2-hydroxyglutarate, and cancer. *Genes Dev*. **27**, 836–852 (2013).
12. Whetstine, J. R. et al. Reversal of histone lysine trimethylation by the JMJD2 family of histone demethylases. *Cell* **125**, 467–481 (2006).
13. Polo, S. E. & Jackson, S. P. Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. *Genes Dev*. **25**, 409–433 (2011).
14. Pierce, A. J., Johnson, R. D., Thompson, L. H. & Jasin, M. XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. *Genes Dev*. **13**, 2633–2638 (1999).
15. Bindra, R. S., Goglia, A. G., Jasin, M. & Powell, S. N. Development of an assay to measure mutagenic non-homologous end-joining repair activity in mammalian cells. *Nucleic Acids Res*. **41**, e115 (2013).
16. Zhou, Y., Caron, P., Legube, G. & Paull, T. T. Quantitation of DNA double-strand break resection intermediates in human cells. *Nucleic Acids Res*. **42**, e19 (2014).
17. Ayrapetov, M. K., Gursoy-Yuzugullu, O., Xu, C., Xu, Y. & Price, B. D. DNA double-strand breaks promote methylation of histone H3 on lysine 9 and transient formation of repressive chromatin. *Proc. Natl Acad. Sci. USA* **111**, 9169–9174 (2014).
18. Sun, Y. et al. Histone H3 methylation links DNA damage detection to activation of the tumour suppressor Tip60. *Nat. Cell Biol*. **11**, 1376–1382 (2009).
19. Williamson, E. A., Wray, J. W., Bansal, P. & Hromas, R. Overview for the histone codes for DNA repair. *Prog. Mol. Biol. Transl*. **110**, 207–227 (2012).
20. Ikura, T. et al. Involvement of the TIP60 histone acetylase complex in DNA repair and apoptosis. *Cell* **102**, 463–473 (2000).
21. Black, J. C. et al. KDM4A lysine demethylase induces site-specific copy gain and rereplication of regions amplified in tumors. *Cell* **154**, 541–555 (2013).
22. Young, L. C., McDonald, D. W. & Hendzel, M. J. Kdm4b histone demethylase is a DNA damage response protein and confers a survival advantage following γ-irradiation. *J. Biol. Chem*. **288**, 21376–21388 (2013).
23. Wang, P. et al. Oncometabolite d-2-hydroxyglutarate inhibits ALKBH DNA repair enzymes and sensitizes *IDH* mutant cells to alkylating agents. *Cell Rep*. **13**, 2353–2361 (2015).
24. Pedersen, M. T. et al. Continual removal of H3K9 promoter methylation by Jmjd2 demethylases is vital for ESC self-renewal and early development. *EMBO J*. **35**, 1550–1564 (2016).
25. Xu, Y. et al. Targeted disruption of ATM leads to growth retardation, chromosomal fragmentation during meiosis, immune defects, and thymic lymphoma. *Genes Dev*. **10**, 2411–2422 (1996).
26. Shan, C.-M. et al. A histone H3K9M mutation traps histone methyltransferase Clr4 to prevent heterochromatin spreading. *eLife* **5**, e17903 (2016).
27. Mallette, F. A. et al. RNF8- and RNF168-dependent degradation of KDM4A/JMJD2A triggers 53BP1 recruitment to DNA damage sites. *EMBO J*. **31**, 1865–1878 (2012).
28. Turcan, S. et al. Mutant-IDH1-dependent chromatin state reprogramming, reversibility, and persistence. *Nat. Genet*. **50**, 62–72 (2018).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

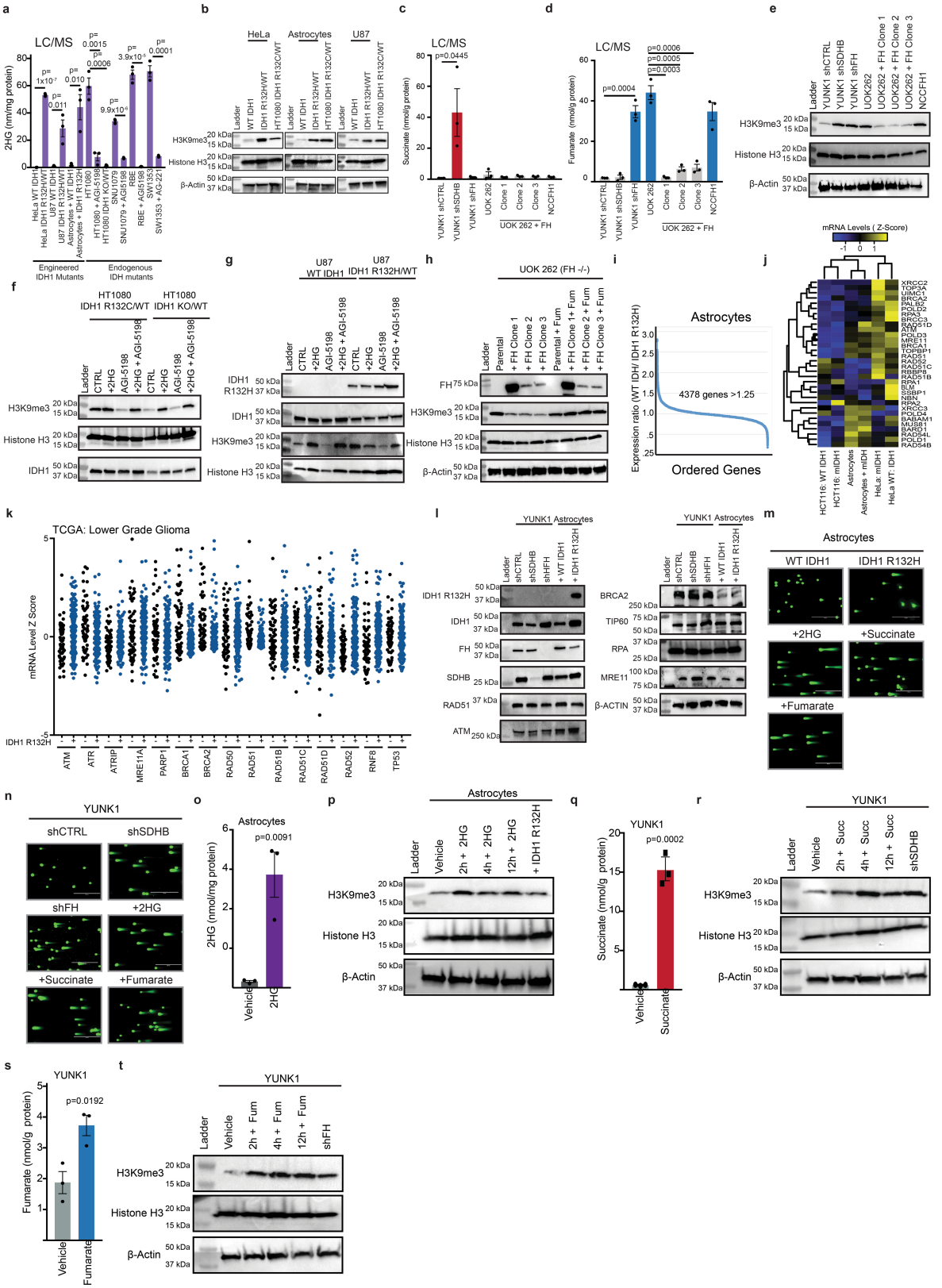
All data generated or analysed during this study are included in this published article (and its supplementary information files, including source data). There are no restrictions on data availability.

Acknowledgements We thank D. Hegan and A. Dhawan for their assistance. This work was supported by the NIH (grants R01ES005775 and R35CA197574 to P.M.G., and R01CA215453 to R.S.B.) and by the American Cancer Society (Research Scholar Grant to R.S.B.). P.L.S. was supported by the NIH National Institute of General Medical Sciences training grant T32GM007223.

Author contributions P.L.S. designed and performed experiments and contributed to all aspects of the study. S.O. designed experiments and performed and analysed imaging studies and laser micro-stripe irradiation. J.D. and N.E. performed DNA repair and cell biology assays and contributed to data analysis and compiling of the manuscript. Y.L. designed and performed the tumour growth delay assays. B.S., X.B. and K.N. provided reagents and contributed to the experiments. L.M. and M.C.K. designed and performed end-resection assays. J.L. performed LC/MS assays. P.M.G. and R.S.B. conceptualized and supervised the study and interpreted the data. P.M.G., R.S.B., M.C.K. and P.L.S. wrote the manuscript.

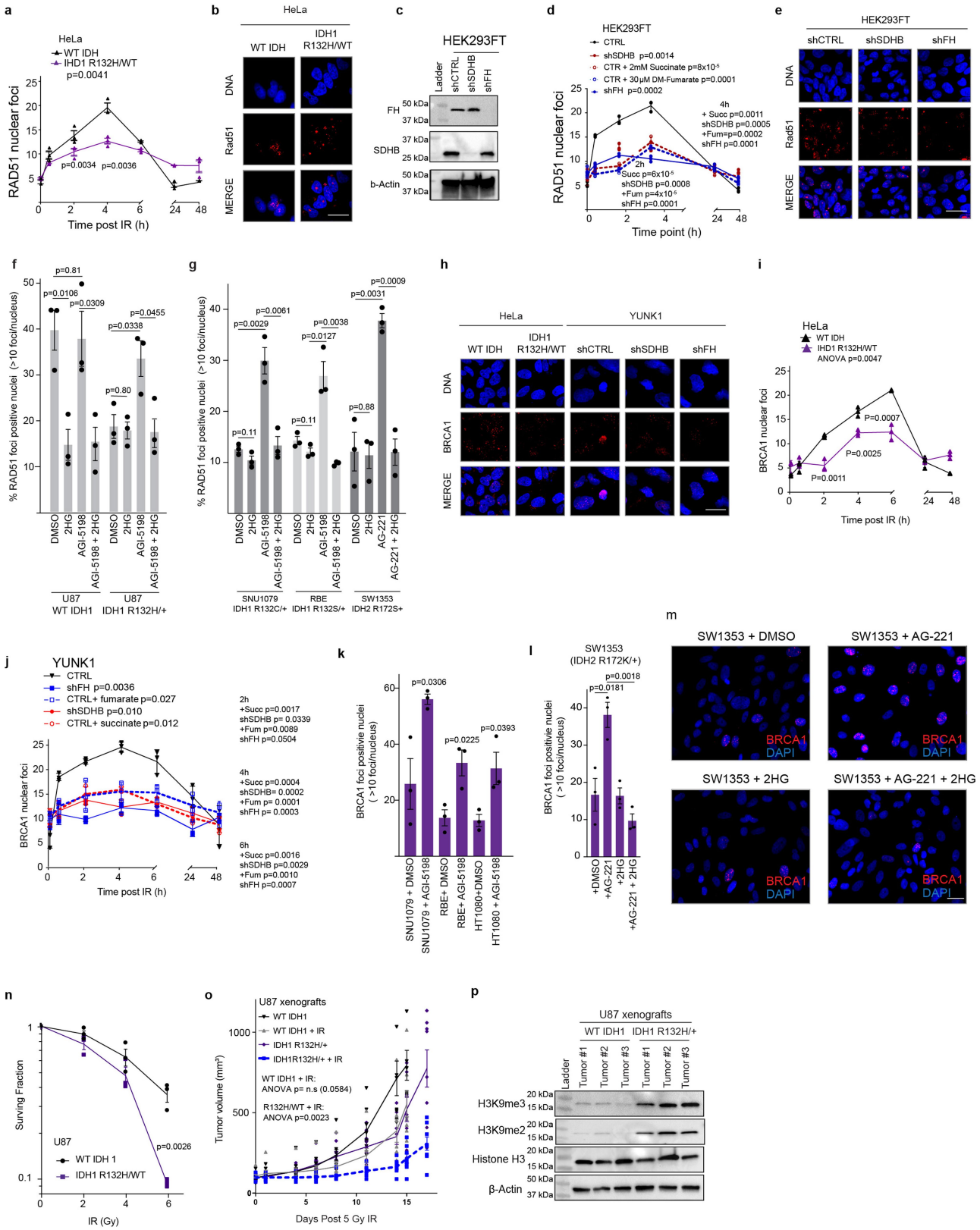
Competing interests R.S.B., P.M.G. and P.L.S. are inventors on US patent application no. 62/344,678 submitted by and assigned to Yale University, which covers compositions and methods for targeting and treating HDR-deficient tumours. P.M.G and R.S.B are founders of and consultants for Cybrexa Therapeutics. P.M.G. is a consultant for pHLIP. Neither company has any conceptual or financial connection to this manuscript.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2363-0.
Correspondence and requests for materials should be addressed to R.S.B. or P.M.G.
Peer review information *Nature* thanks Brendan Price, Yue Xiong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



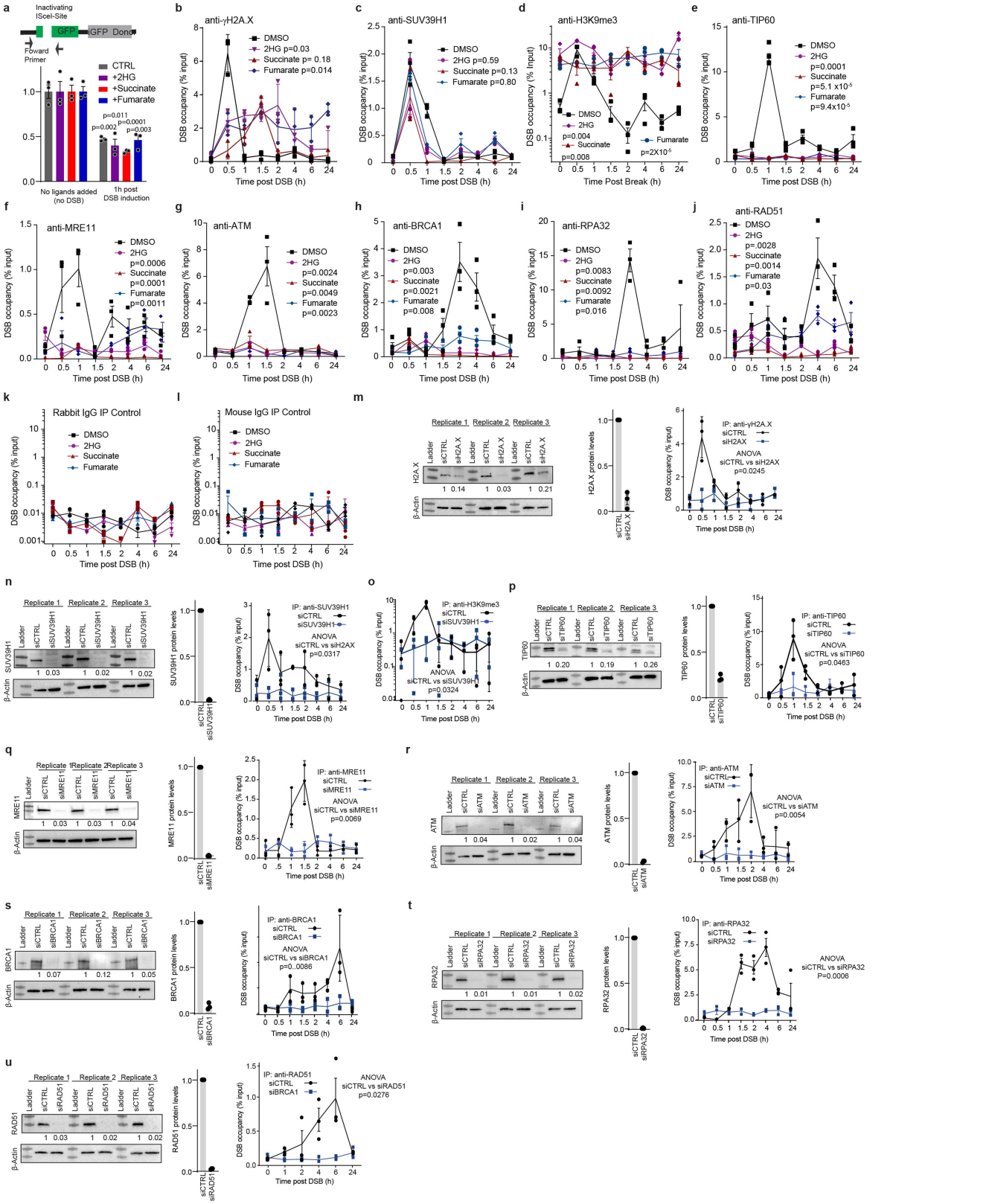
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Elevated oncometabolites induce H3K9me3 and influence global gene expression, but do not alter expression of HDR genes. a, Liquid chromatography with mass spectrometry (LC–MS) quantification of 2HG levels in cells with engineered *IDH1* mutations and in cells with endogenous *IDH1* and *IDH2* mutations, as indicated. Note that engineered *IDH1*-mutant cells produce 2HG at levels similar to the endogenous mutant cell lines, and that the HT1080 cells with CRISPR–Cas9 knockout of the mutant $IDH1^{R132C}$ allele ($IDH1^{KO/+}$) show reduction of 2HG production to levels similar to those seen in IDH1 wild-type cells. Levels of 2HG are also quantified in endogenous mutant IDH1 or IDH2 cells after treatment with the mutant-IDH1-specific inhibitor AGI-5198 and the mutant-IDH2-specific inhibitor AG-221. **b**, Western blot analysis of H3K9me3 levels in cell lines with WT IDH1 or engineered to express mutant IDH1, as indicated, compared with the HT1080 cell line with an endogenous $IDH1^{R132C}$ mutation. **c, d**, LC–MS analysis of succinate (**c**) and fumarate (**d**) in: YUNK1 cells with shRNA suppression of *SDHB* or *FH* compared with shCTRL (non-targeting) control cells, the endogenous $FH^{-/-}$ UOK 262 renal cell carcinoma cell line with and without *FH* cDNA complementation (clones 1–3), and in the $FH^{-/-}$ NCCFH1 renal cell carcinoma cell line. **e**, Western blot analysis of H3K9me3 in YUNK1 cells with shRNA suppression of *SDHB* or *FH* compared to shCTRL (non-targeting) control cells, the endogenous $FH^{-/-}$ UOK 262 renal cell carcinoma cell line with and without *FH* cDNA complementation, and in the $FH^{-/-}$ NCCFH1 renal cell carcinoma cell line. This experiment was repeated twice with similar results. **f**, Western blot analysis of H3K9me3 in HT1080 cells ($IDH1^{R132C/+}$) and in HT1080 cells with CRISPR–Cas9 knockout of the mutant *IDH1* allele ($IDH1^{KO/+}$), and with or without treatment with 500 μM 2HG, 1 mM AGI-5198 or 1 mM AGI-5198 and 500 μM 2HG. This experiment was repeated twice with similar results. **g**, Western blot analysis of H3K9me3 in WT IDH1 U87 glioblastoma cells and in U87 cells with CRISPR–Cas9 knock-in of the mutant $IDH1^{R132H}$ allele and with or without treatment with 500 μM 2HG, 1 μM AGI-5198 or 1 μM AGI-5198 and 500 μM 2HG. This experiment was repeated three times with similar results. **h**, Western blot analysis of H3K9me3 in $FH^{-/-}$ UOK 262 cells with or without *FH* cDNA complementation (clones 1–3) and with or without treatment with 30 μM dimethyl fumarate. This experiment was repeated twice with similar results. **i**, Plot of microarray gene expression analysis of astrocytes either overexpressing $IDH1^{WT}$ or $IDH1^{R132H}$. Genes were plotted in descending order based on the ratio of $IDH1^{WT}$ expression levels to that of $IDH1^{R132H}$. Genes with a high value are highly expressed in IDH^{WT} cells compared to $IDH1^{R13H}$ cells, and therefore represent genes that are putatively suppressed in $IDH1^{R132H}$-expressing cells. **j**, Heat map representation of microarray expression analysis of HDR-associated genes in matched pairs of otherwise isogenic IDH^{WT} and $IDH1^{R132H}$ mutant cells: IDH^{WT} and $IDH1^{R132H/+}$ HCT116 cells; human immortalized astrocytes expressing IDH^{WT} or $IDH1^{R132H}$, and IDH^{WT} and $IDH1^{R132H/+}$ HeLa cells. **k**, Dot plot of expression levels of HDR-associated genes from the TCGA lower-grade glioma mRNA-seq dataset. For each gene, patient samples are separated by IDH status. **l**, Western blot analyses of IDH1(R132H), total IDH1, FH, SDHB, RAD51, ATM, BRCA2, TIP60, RPA and MRE11 in YUNK1 cells with shRNA suppression of *SDHB* or *FH* compared with shCTRL (non-targeting) control cells and in astrocytes overexpressing IDH1(WT) or IDH1(R132H). This experiment was repeated four times with similar results. **m**, **n**, Representative images of neutral comet assays performed in immortalized astrocytes overexpressing IDH1(WT) or IDH1(R132H) or treated with 2HG, 2 mM succinate or 30 μM dimethyl fumarate (**m**) and in YUNK1 cells after shRNA suppression of *FH* or *SDHB* or addition of 500 μM octyl-(R)-2HG, 2 mM succinate or 30 μM dimethyl fumarate (**n**). Scale bars, 400 μm. **o**, LC–MS quantification of 2HG in astrocytes after 2 h treatment with 500 μM R-octyl-2HG. **p**, Western blot analysis of H3K9me3 levels in astrocytes treated with 500 μM R-octyl-2HG for 2, 4 or 12 h, compared to vehicle control (DMSO cells) or astrocytes expressing IDH1(R132H). This experiment was repeated two times with similar results. **q**, LC–MS quantification of succinate in YUNK1 cells after 2 h treatment with 2 mM succinate. **r**, Western blot analysis of H3K9me3 levels in YUNK1 cells treated with 2 mM succinate for 2, 4 or 12 h, compared to vehicle control (DMSO) cells or YUNK1 cells with shRNA suppression of *SDHB*. This experiment was repeated two times with similar results. **s**, LC–MS quantification of fumarate in YUNK1 cells after 2 h treatment with 30 μM dimethyl fumarate. **t**, Western blot analysis of H3K9me3 levels in YUNK1 cells treated with 2 mM succinate for 2, 4 or 12 h, compared to vehicle control (DMSO) cells or YUNK1 cells with shRNA suppression of *FH*. This experiment was repeated two times with similar results. In **a**, **c**, **d, o, q** and **s**, data are mean ± s.e.m. with $n = 3$ biological replicates; statistical analysis is by two-tailed unpaired *t*-test; df = 4, with *P* values as indicated.



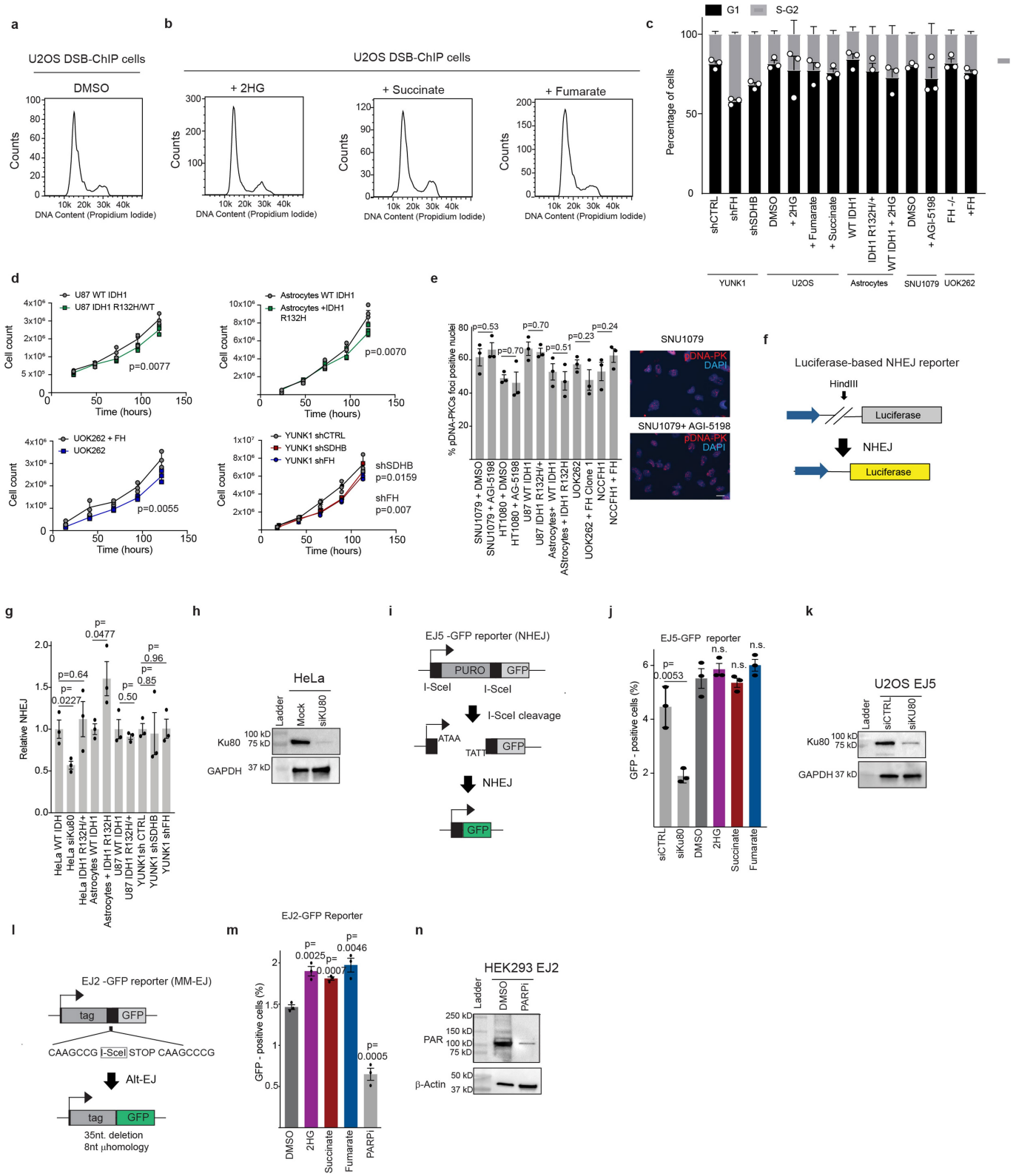
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Oncometabolites directly impair formation of RAD51 and BRCA1 foci after ionizing radiation and cause radiosensitization. a, Quantification of RAD51 nuclear foci at the indicated time points after 2 Gy ionizing radiation (IR), in IDH^{WT} or $IDH1^{R132H/+}$ HeLa cells. Statistical analysis of time course by ANOVA, $F = 35.15$, df = 1. Statistical analysis at specific time points by two-tailed unpaired *t*-test, df = 4. **b**, Representative images of RAD51 nuclear foci 4 h after 2 Gy ionizing radiation in IDH^{WT} or $IDH1^{R132H/+}$ HeLa cells. These experiments were performed three times as quantified in **a**. Scale bar, 10 µm. **c**, Western blot analysis of FH and SDHB in HEK293FT cells with shCTRL (non-targeting control) or shRNA suppression of *SDHB* or *FH*. These experiments were repeated six times. **d**, Quantification of RAD51 nuclear foci at the indicated time points following 2 Gy ionizing radiation in HEK293FT cells with shRNA suppression of *SDHB* or *FH* compared to non-targeting control (shCTRL) cells, and in shCTRL cells treated with either 2 mM succinate or 30 µM dimethyl fumarate. Statistical analysis of time course by ANOVA. *F* values versus control: shSDHB, $F = 61.55$; shFH, $F = 195.3$; +succinate, $F = 258.2$; +fumarate, $F = 242.4$; df = 1 for all tests. Statistical analysis at indicated time points by two-tailed unpaired, *t*-test; df = 4. **e**, Representative images of RAD51 nuclear foci at 4 h post ionizing radiation in HEK293FT cells with shRNA suppression of *SDHB* or *FH* compared to non-targeting control (shCTRL) cells. Scale bar, 10 µm. These experiments were performed three times and are quantified in **d**. (**f**) Quantification of cells with RAD51 foci-positive nuclei (>10 foci per nucleus) 4 h after 2 Gy ionizing radiation treatment of parental U87 glioma cells ($IDH1^{WT}$) and in U87 cells with CRISPR–Cas9-mediated knock-in of an $IDH1^{R132H}$ allele at the endogenous locus ($IDH1^{R132H/+}$). Cells were treated as indicated with or without 500 µM octyl-(R)-2HG (2HG) for 24 h, 1 mM AG1-5198 for 5 days or a combination thereof, before irradiation. **g**, Quantification of RAD51 foci-positive nuclei in SNU1079 ($IDH1^{R132C/+}$) and RBE ($IDH1^{R132S/+}$) cholangiocarcinoma cells, and SW1353 ($IDH2^{R172K/+}$) chondrosarcoma cells 4 h after 2 Gy ionizing radiation. Mutant IDH1 cells were treated with or without 2HG, AGI-5198 or AGI-5198 + 2HG as in **f** and the *IDH2*-mutant SW1353 cells were treated with or without 2HG or AG-221 or AG-221 + 2HG. **h**, Representative images of BRCA1 nuclear foci at 4 h after 2 Gy ionizing radiation. Scale bar, 10 µm. This experiment was repeated three times and is quantified in **i** and **j**. **i**, Quantification of BRCA1 nuclear foci at the indicated time points after 2 Gy ionizing radiation, in IDH^{WT} or $IDH1^{R132H/+}$ HeLa cells. Statistical analysis of time course by ANOVA; $F = 41.35$, df = 1. Statistical analysis at specific time points by two-tailed unpaired *t*-test; df = 4. **j**, Quantification of BRCA1 nuclear foci at the indicated time points following 2 Gy ionizing radiation in YUNK1 cells with shRNA suppression of *SDHB* or *FH* compared to non-targeting control (shCTRL) cells, and in shCTRL cells treated with either 2 mM succinate or 30 µM dimethyl fumarate. Statistical analysis of time course by ANOVA. *F* values versus control: shSDHB, $F = 20.92$; shFH, $F = 37.42$; +succinate, $F = 19.89$; +fumarate $F = 11.57$, df = 1 for all tests. Statistical analysis at indicated time points by two-tailed unpaired, *t*-test, df = 4. **k**, Quantification of BRCA1 foci-positive cells (>10 nuclear BRCA1 foci) 4 h after 2 Gy ionizing radiation in SNU1079 ($IDH1^{R132C/+}$), RBE ($IDH1^{R132S/+}$) and HT1080 ($IDH1^{R132C/+}$) cells treated with 1 µM AGI-5198 or DMSO control. **l**, **m**, Quantification (**l**) and representative images (**m**) of BRCA1 nuclear foci 4 h after 2 Gy ionizing radiation in SW1353 ($IDH2^{R172K/+}$) cells treated as indicated with DMSO (control), 500 µM 2HG, 5 µM AG-221 or 5 µM AG-221 and 500 µM 2HG. Scale bar, 10 µm. Images in **m** are from experiments that were repeated three times and are quantified in **l**. **n**, Clonogenic survival assay in U87 glioblastoma cells with or without mutant IDH1(R132H) treated with the indicated doses of ionizing radiation. **o**, Tumour growth delay assay in $IDH1^{WT}$ and $IDH1^{R132H/+}$ U87 tumour xenografts. Mice were treated with a single dose of 8 Gy ionizing radiation or mock irradiated at a tumour volume of 150 mm^3. $N = 8$ mice per group each with a single tumour. Statistical analysis by ANOVA ($IDH1^{WT}$, $F = 4.25$, df = 1; $IDH1^{R132H}$, $F = 13.76$, df = 1). **p**, Western blot analysis of H3K9me3 in U87 $IDH1^{WT}$ and U87 $IDH1^{R132H/+}$ tumour xenografts. This experiment was performed three times with similar results. In **f**, **g**, **k**, **l**, data are mean ± s.e.m. with $n = 3$ biological replicates; statistical analysis by two-tailed unpaired *t*-test; df = 4. In **a**, **d**, **i**, **j**, **n**, data are mean ± s.e.m. of 3 biological replicates; statistical analysis by ANOVA: *P* values are indicated.



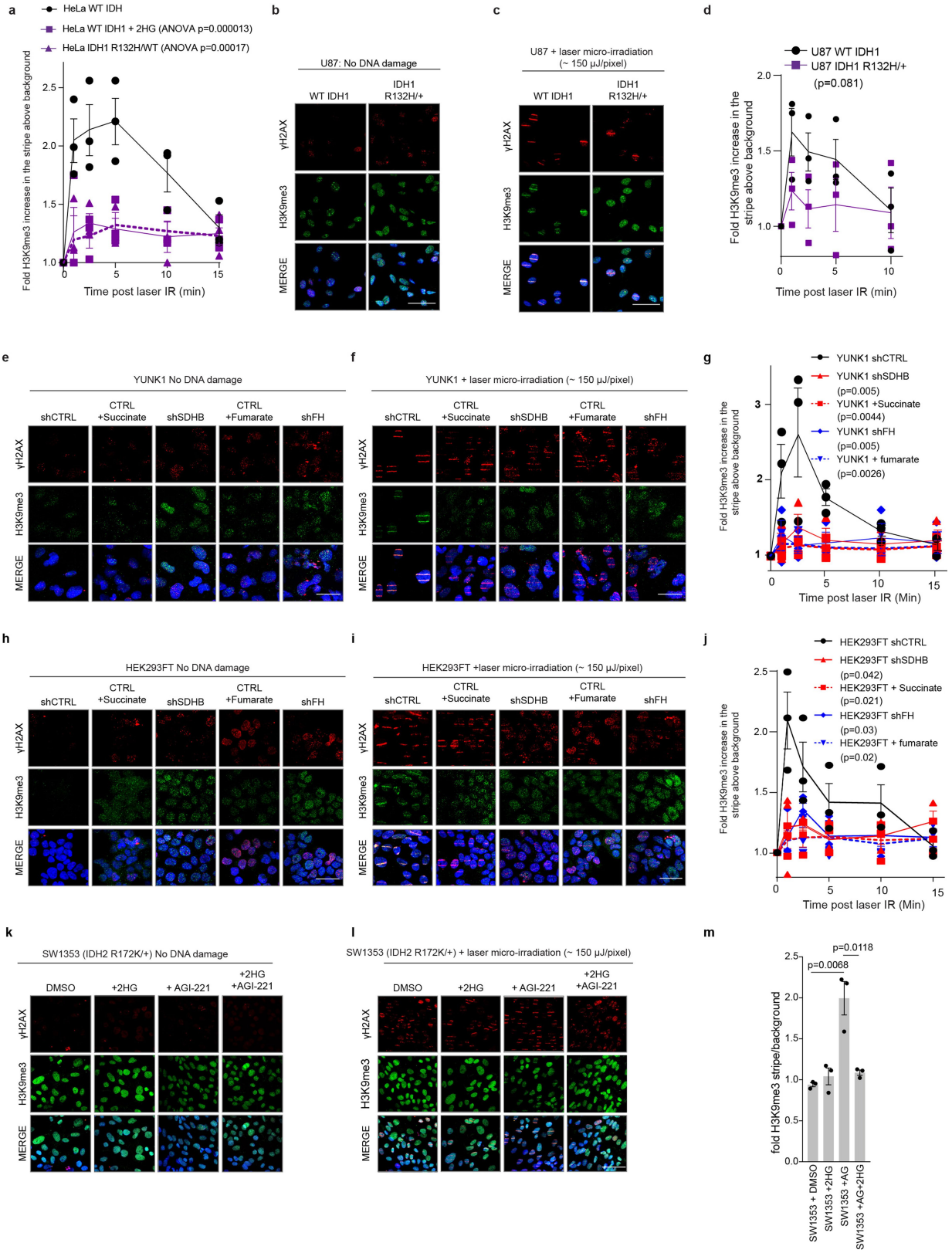
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Supporting data for DSB–ChIP assays to analyse recruitment of DNA repair factors to an I-SceI-induced, site-specific DNA DSB. a, Top, schematic of the qPCR analysis to assess I-SceI cleavage at the DSB–ChIP site. Bottom, qPCR analysis to assay site-specific DNA DSB induction 1 h after addition of the ligands Shield-1 and triamcinolone. Reduced amplification of the genomic DNA across the I-SceI target indicates that cleavage at that site has occurred; short bars indicate reduced PCR amplification and therefore successful cleavage. **b**–**j**, Pertaining to Fig. 2b, line graphs of percent input values for DSB–ChIP assays performed with the indicated antibodies in cells treated as indicated with either DMSO, 500 μM ocyl-R-2HG, 2 mM succinate or 30 μM dimethyl fumarate, for the following factors: γH2A.X (**b**; +2HG, $F = 10.8$; +succinate, $F = 2.682$; +fumarate, $F = 17.8$); SUV39H1 (**c**; +2HG, $F = 0.33$; +succinate, $F = 3.5$; +fumarate, $F = 0.07$); H3K9me3 (**d**; +2HG, $F = 124.4$; +succinate, $F = 25.21$; +fumarate, $F = 517$); TIP60 (**e**; +2HG, $F = 218$; +succinate, $F = 340.7$; +fumarate, $F = 248.6$); MRE11 (**f**; +2HG, $F = 97.7$; +succinate, $F = 209.9$; +fumarate, $F = 71.2$); ATM (**g**; +2HG, $F = 46.8$; +succinate, $F = 31.7$; +fumarate, $F = 47.3$); BRCA1 (**h**; +2HG, $F = 50.0$; +succinate, $F = 50.7$; +fumarate, $F = 24.1$); RPA32 (**i**; +2HG, $F = 23.59$; +succinate, $F = 22.3$; +fumarate, $F = 16.0$); and RAD51 (**j**; +2HG, $F = 43.3$; +succinate, $F = 61.8$; +fumarate, $F = 11.75$) at the indicated time points after addition of triamcinolone and Shield-1 to induce an I-SceI break in DSB–ChIP U2OS cells. **k**, **l**, Line graphs of per cent input values for DSB–ChIP assays performed with IgG controls for rabbit IgG (**k**) and mouse IgG (**l**) at the indicated time points post addition of triamcinolone and Shield-1 to induce an I-SceI break in DSB–ChIP cells treated with either DMSO, 500 μM octyl-R-2HG, 2 mM succinate or 30 μM dimethyl fumarate. **m**–**u**, Antibody validation experiments for the DSB–ChIP assays. Each panel includes a western blot analysis of target-protein knockdown by siRNA (with three biological replicates in each case and quantification of western blot band intensities normalized to β-actin loading control and presented below each lane), an accompanying bar graph quantifying the knockdown data showing mean ± s.e.m., with dots indicating individual values for each of the biological replicates, and then a DSB–ChIP assay performed with the same antibody for the respective target protein with and without siRNA knockdown, as follows: γH2A.X (**m**; $F = 12.3$, df = 1); SUV39H1 (**n**; $F = 10.49$, df = 1); H3K9me3 ChIP assay after siRNA knockdown of the H3K9 methyltransferase SUV39H1 (as shown in **n**) (**o**; $F = 10.34$, df = 1); TIP60 (**p**; $F = 8.13$, df = 1); MRE11 (**q**; $F = 26.1$, df = 1); ATM (**r**; $F = 29.9$, df = 1); BRCA1 (**s**; $F = 23.17$, df = 1), RPA32 (**t**; $F = 97.5$, df = 1) and RAD51 (**u**; $F = 9.4$, df = 1). In **a**, data are mean ± s.e.m. with $n = 3$ biological replicates; statistical analysis by two-tailed unpaired *t*-test; df = 4, *P* values as indicated. In **b**–**u**, lines run through the mean ± s.e.m. of three biological replicates; statistical analysis by ANOVA; *P* values indicated.



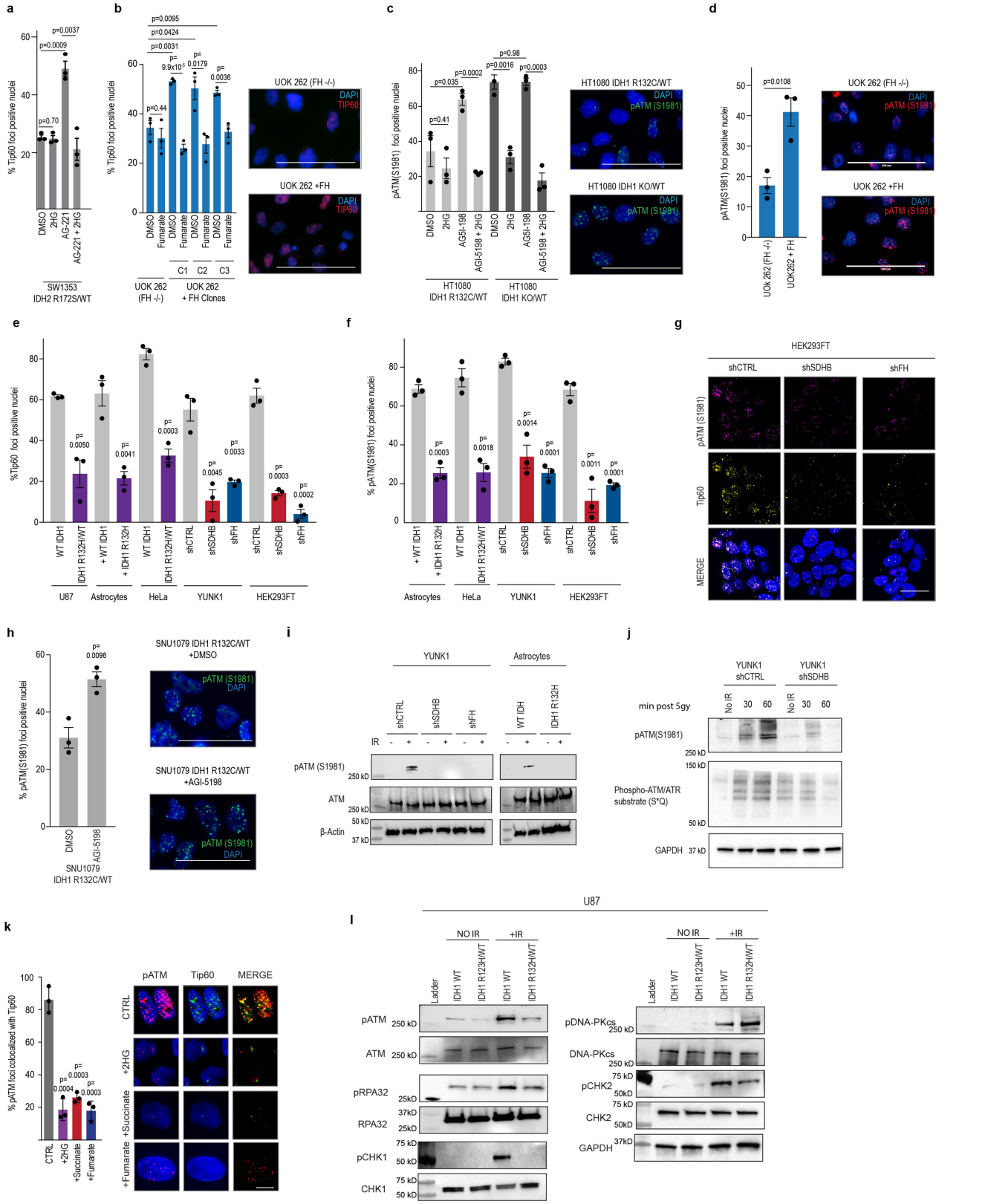
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Elevated metabolites have minimal effects on cell cycle distribution, growth rates and NHEJ DNA repair. a, **b**, Cell cycle profile plots based on DNA content by flow cytometry for U2OS DSB–ChIP cells treated with either DMSO (control) (**a**) or with the indicated metabolites: 500 μM octyl-R-2HG, 2 mM succinate, or 30 μM dimethyl fumarate (**b**). **c**, Quantification of cell cycle analyses in: YUNK1 cells with shRNA suppression of *SDHB* or *FH* compared to shCTRL (non-targeting) controls; U2OS DSB–ChIP cells treated with DMSO, 500 μM octyl-R-2HG, 2 mM succinate, or 30 μM dimethyl fumarate; astrocytes expressing IDH1(WT) or IDH1(R132H); SNU1079 ($IDH1^{R132C/+}$) cells treated with DMSO or 1 μM AGI-5198; and UOK 262 $FH^{-/-}$ renal cell carcinoma cells with and without *FH* cDNA complementation. **d**, Serial cell counts over time of the indicated cells in standard culture conditions: U87 $IDH1^{R132H/+}$ glioblastoma cells compared to $IDH1^{WT}$ U87 cells; astrocytes expressing IDH1(WT) or IDH1(R132H); UOK 262 $FH^{-/-}$ renal cell carcinoma cells with and without *FH* cDNA complementation; and YUNK1 cells with shRNA suppression of *SDHB* or *FH* compared to shCTRL (non-targeting) control cells. Data are mean ± s.e.m. of three biological replicates; *P* values by ANOVA are indicated. Additional statistics: for U87, $F = 24.69$ (df = 1); astrocytes, $F = 26.00$ (df = 1); UOK262, $F = 29.64$ (df = 1); YUNK1 shSDHB, $F = 16.11$ (df = 1), and YUNK1 shFH, $F = 89.47$ (df = 1). **e**, pDNA-PKcs foci formation (using an antibody to phosphorylated DNA-PKcs) at 4 h after 2 Gy ionizing radiation quantified as percentage foci-positive nuclei (>10 foci per nucleus) in: SNU1079 cells ($IDH1^{R132H/+}$) and HT1080 ($IDH1^{R132C/+}$) cells treated with 1 μM AGI-5198 or DMSO control; U87 glioblastoma $IDH1^{WT}$ and $IDH1^{R132H/+}$ cells; astrocytes overexpressing IDH1(WT) or IDH1(R132H); UOK 262 $FH^{-/-}$ renal cell carcinoma cells with and without *FH* cDNA complementation; and NCCFH1 $FH^{-/-}$ renal cell carcinoma cells with or without *FH* complementation. Representative images from SNU1079 ($IDH1^{R132C/+}$) cells with and without AGI-5198 treatment are shown. Scale bar, 20 μm. **f**, Schematic of the luciferase based NHEJ assay. **g**, Quantification of relative NHEJ by the luciferase-based NHEJ assay in HeLa cells, Hela cells after siRNA suppression of *KU80* (also known as *XRCC5*, positive control for NHEJ deficiency), $IDH1^{R132H/+}$ HeLa cells, astrocytes overexpressing IDH1(WT) or IDH1(R132H), U87 glioblastoma $IDH1^{WT}$ and U87 $IDH1^{R132H/+}$ cells, and YUNK1 cells with shRNA suppression of *SDHB* or *FH*, compared to shCTRL (non-targeting) control cells. **h**, Western blot analysis to confirm KU80 knockdown after siRNA suppression of *KU80* in HeLa cells. This experiment was performed two times with similar results. **i**, Schematic of the EJ5–GFP NHEJ reporter. **j**, Quantification of NHEJ using the EJ5 chromosomally integrated reporter assay in U2OS-EJ5 reporter cells after treatment of cells with 500 μM octyl-R-2HG, 2 mM succinate or 30 μM dimethyl fumarate, compared to DMSO control and to cells with siRNA knockdown of *KU80*. **k**, Quantification of KU80 levels after siRNA suppression of *KU80* in U2OS-EJ5 reporter cells. This experiment was performed two times with similar results. **l**, Schematic diagram of the EJ2–GFP reporter to assay MMEJ. **m**, Quantification of EJ2 MMEJ reporter activity in EJ2-HEK293FT cells after treatment with 500 μM octyl-R-2HG, 2 mM succinate or 30 μM dimethyl fumarate compared to DMSO control. Treatment with PARP inhibitor (PARPi) BMN-673 is used as a positive control for PARP-dependent MMEJ. **n**, Western blot levels of poly-ADP-ribose (PAR) polymers and poly-ribosylated proteins levels after treatment with 10 nM of the PARPi BMN-673. This experiment was performed 2 times with similar results. In **c**, **e**, **g**, **j**, **m**, data are mean ± s.e.m. with $n = 3$ biological replicates; statistical analysis by two-tailed unpaired *t*-test; df = 4.



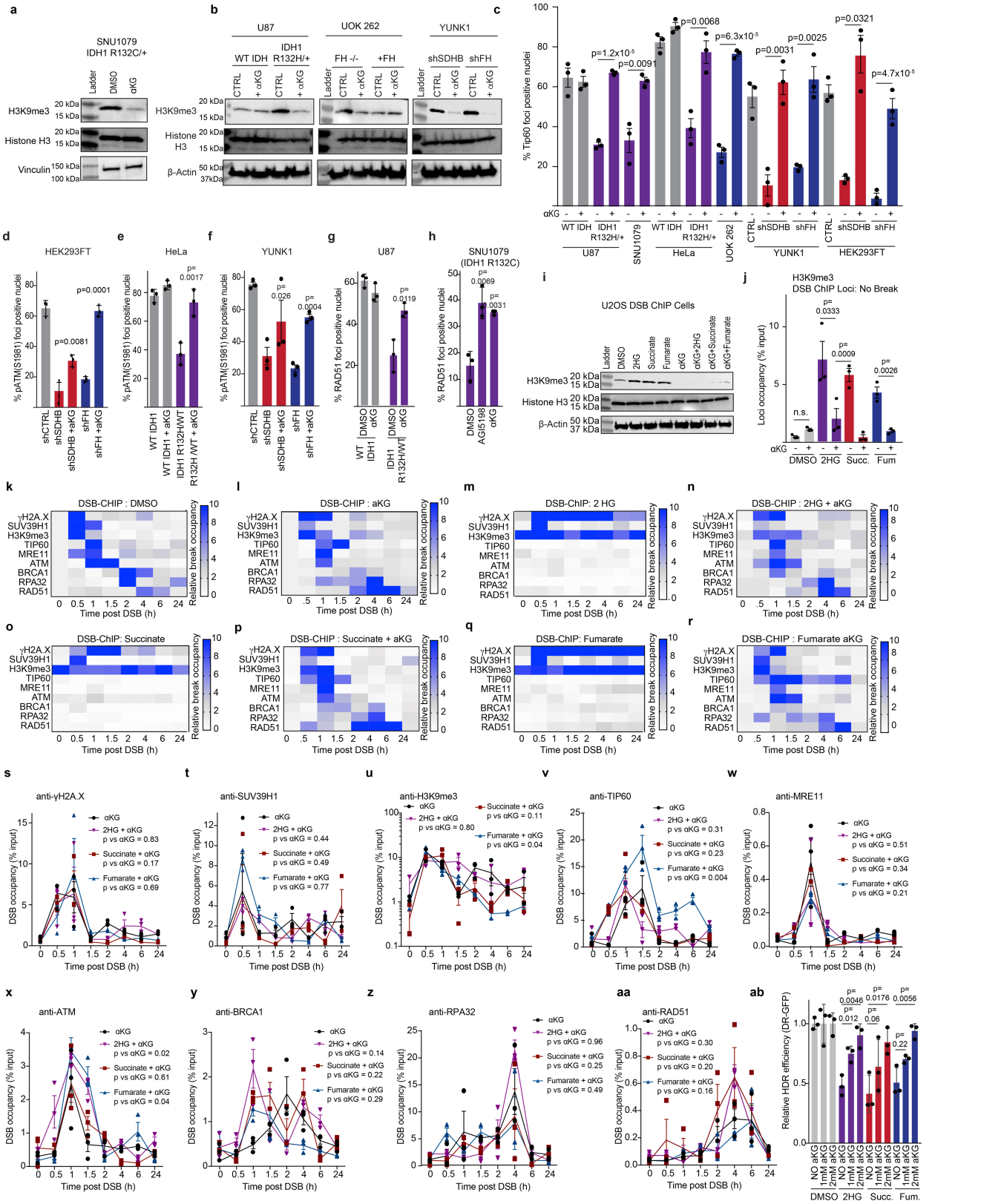
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Elevated metabolites disrupt normal H3K9me3 deposition at sites of DNA damage caused by laser-stripe micro-irradiation. a, Quantification of H3K9me3 intensity in the laser micro-irradiated stripe above background after laser micro-irradiation induction of DNA damage in HeLa cells treated with or without 2HG or expressing IDH1(R132H), as indicated. ($F = 49.27$, df = 1). **b**, **c**, Representative immunofluorescence of γH2AX and H3K9me3 in cell nuclei in parental U87 glioblastoma cells (WT IDH1) and U87 cells with CRISPR–Cas9-mediated knock-in of an $IDH1^{R132H}$ allele at the endogenous locus ($IDH1^{R132H/+}$) without DNA damage (**b**) or 1 min after laser micro-irradiation induction of DNA damage at 150 µJ per pixel (**c**). Scale bars, 20 µm. **d**, Quantification of H3K9me3 intensity in the laser micro-irradiated stripe above background after the indicated time points in the U87 cell line matched pair after laser micro-irradiation. ($F = 5.360$, df = 1). **e**, **f**, Representative images of undamaged (**e**) and laser micro-stripe irradiated (**f**) YUNK1 cells with shRNA suppression of *SDHB* or *FH* compared to non-targeting control shRNA (shCTRL), or in YUNK1 cells treated with 2 mM succinate or 30 µM dimethyl fumarate. Cells were pretreated with exogenous metabolites 24 h before micro-stripe irradiation and then were analysed 1 min after irradiation. Scale bars, 20 µm. **g**, Quantification of H3K9me3 intensity in the laser micro-irradiated stripes above background at the indicated time points after laser micro-irradiation in the YUNK1 cells treated as indicated. (shSDHB, $F = 31.25$, df = 1; +succinate, $F = 33.80$, df = 1; shFH, $F = 32.25$, df = 1; +fumarate, $F = 44.39$, df = 1). **h, i**, Representative images of undamaged (**h**) and laser micro-stripe irradiated (**i**) HEK293FT cells with shRNA suppression of *SDHB* or *FH* compared to non-target control shRNA (shCTRL), or in HEK293FT cells treated with 2 mM succinate or 30 µM dimethyl fumarate. Cells were pretreated with exogenous metabolites 24 h before micro-stripe irradiation and were analysed at 1 min after irradiation. Scale bars, 20 µm. **j**, Quantification of H3K9me3 intensity in the laser micro-irradiated stripe above background at the indicated times after laser micro-irradiation in the HEK293FT cells treated as indicated. (shSDHB, $F = 8.68$, df = 1; +succinate, $F = 13.75$, df = 1; shFH, $F = 10.84$, df = 1; +fumarate $F = 14.05$, df = 1). **k, l**, Representative images of undamaged (**k**) and laser micro-stripe irradiated (**l**) SW1353 ($IDH2^{R172K/+}$) chondrosarcoma cells treated as indicated with DMSO, 500 µM octyl-(R)-2HG, 5 µM AG-221 or 5 µM AG-221 and 500 µM octyl-(R)-2HG. Scale bars, 20 µm. **m**, Quantification of H3K9me3 intensity in the laser micro-irradiated stripe above background at 2 min after laser micro-irradiation in the SW1353 ($IDH1^{R172K/+}$) chondrosarcoma cells treated as indicated. In **a**, **d**, **g**, **j**, line runs through the mean ± s.e.m. with $n = 3$ biological replicates for each time point; statistical analysis by ANOVA. In **m**, data are mean ± s.e.m. with $n = 3$ biological replicates; statistical analysis by two-tailed unpaired *t*-test; df = 4, *P* values are indicated.



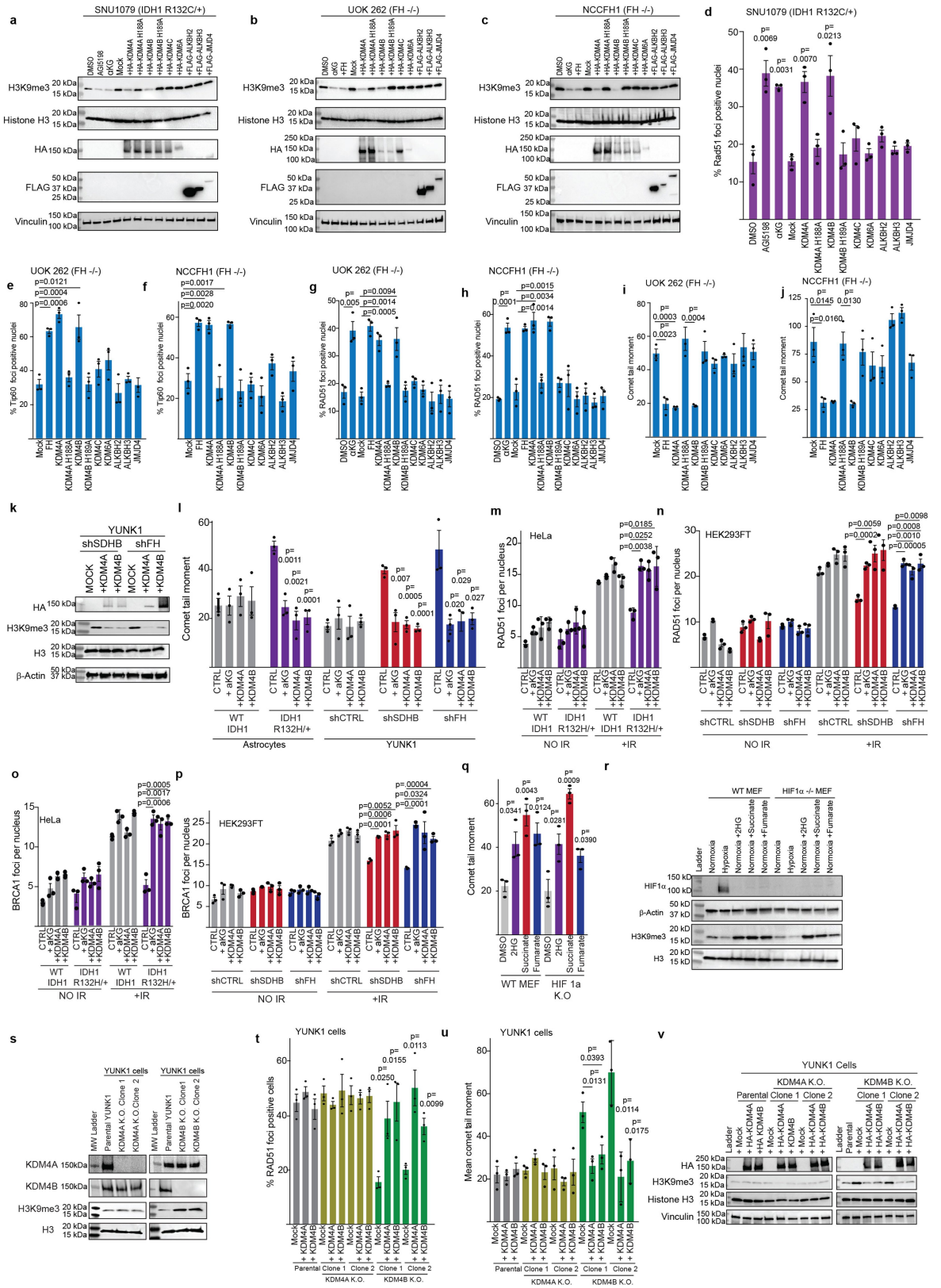
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Oncometabolites impair TIP60 and ATM recruitment and activation following treatment of cells with ionizing radiation. **a**, Quantification of TIP60 foci-positive nuclei 1 h after 2 Gy ionizing radiation in SW1353 (*IDH2*$^{R172K/+}$) cells treated as indicated with DMSO (control), 500 μM 2HG, 5 μM AG-221 or 5 μM AG-221 and 500 μM 2HG. **b**, Quantification and representative images of TIP60 foci-positive nuclei 4 h after 2 Gy ionizing radiation in FH-deficient UOK 262 renal cell carcinoma cells (*FH*$^{-/-}$) and in three subclones complemented with *FH* cDNA, with or without treatment with 30 μM dimethyl fumarate for 24 h before 2 Gy ionizing radiation. Scale bars, 100 μm. **c**, Quantification and representative images of phosphorylated ATM on residue S1981 (pATM S1981) foci-positive nuclei 4 h after 2 Gy ionizing radiation in HT1080 fibrosarcoma cells (*IDH1*$^{R132C/+}$) and in HT1080 cells with CRISPR–Cas9-mediated knockout of the *IDH1*^{R132C} allele (*IDH1*$^{KO/+}$) treated as indicated with DMSO control, 2HG, AGI-5198 or AGI-5198 + 2HG. Scale bars, 100 μm. **d**, Quantification and representative images of pATM (S1981) foci-positive nuclei 4 h after 2 Gy ionizing radiation in *FH*-deficient UOK 262 renal cell carcinoma cells (*FH*$^{-/-}$) and in a subclone complemented with *FH* cDNA. Scale bars, 100 μm. **e**, Quantification of cells with TIP60 foci-positive nuclei (>10 foci per nucleus) 1 h after 2 Gy ionizing radiation in U87 *IDH1*^{WT} and U87 *IDH1*$^{R132H/+}$ glioblastoma cells, immortalized astrocytes overexpressing IDH1(WT) or IDH1(R132H), *IDH1*^{WT} and *IDH1*$^{R132H/+}$ HeLa cells, and YUNK1 and HEK293FT cells with shRNA suppression of *SDHB* (shSDHB) or *FH* (shFH), compared to non-targeting control shRNA (shCTRL). **f**, Quantification of cells with pATM S1981 foci-positive nuclei (>10 foci per nucleus) 1 h after 2 Gy ionizing radiation in immortalized astrocytes overexpressing IDH1(WT) or IDH1(R132H), *IDH1*^{WT} HeLa cells, *IDH1*$^{R132H/+}$ HeLa cells, and YUNK1 and HEK293FT cells with shRNA suppression of *SDHB* (shSDHB) or *FH* (shFH) compared to non-targeting control shRNA (shCTRL). **g**, Representative immunofluorescent of TIP60 and pATM S1981 nuclear foci 1 h after 2 Gy ionizing radiation in HEK293FT cells with shRNA suppression of *SDHB* or *FH* compared to non-targeting control shRNA (shCTRL). Scale bar, 20 μm. **h**, Quantification of cells with pATM S1981 foci-positive nuclei (>10 foci per nucleus) and representative immunofluorescent images of pATM S1981 foci 1 h after 2 Gy ionizing radiation in SNU1079 (*IDH1*$^{R132C/+}$) cholangiocarcinoma cells treated with IDH1 inhibitor (1 μM AGI-5198) or DMSO control. Scale bars, 100 μm. **i**, Western blot analysis of pATM S981 and total ATM 1 h after 2 Gy ionizing radiation in YUNK1 cells with shRNA suppression of *SDHB* (shSDHB) or *FH* (shFH) compared to non-targeting control shRNA (shCTRL) and in immortalized astrocytes overexpressing IDH1(WT) or IDH1(R132H). This experiment was performed three times with similar results. **j**, Western blot analysis of phospho-ATM S1981 (pATM) and phospho-S*Q motifs in YUNK1 cells with shRNA suppression of *SDHB* (shSDHB) as compared to a non-targeting control shRNA (shCTRL). This experiment was performed two times with similar results. **k**, Quantification and representative images of pATM and TIP60 foci colocalization in HeLa cells treated with DMSO control (CTRL), 500 μM octyl-(R)-2HG, 2 mM succinate or 30 μM dimethyl fumarate. Scale bar, 5 μm. **l**, Western blot analysis of pATM S1981 and total ATM, phospho-ATR and total ATR, phospho-RPA32 and total RPA32, phospho-CHK1 and total CHK1, phospho-DNA-PKcs and total DNA-PKcs, and phospho-CHK2 and total CHK2 in *IDH1*^{WT} or *IDH1*$^{R132H/+}$ U87 glioblastoma cells 1 h after 5 Gy ionizing radiation, compared to unirradiated controls. This experiment was performed two times with similar results. For **a**–**f**, **h**, **k**, data are mean ± s.e.m. with $n = 3$ biological replicates; statistical analysis by two-tailed unpaired *t*-test; df = 4, *P* values as indicated.



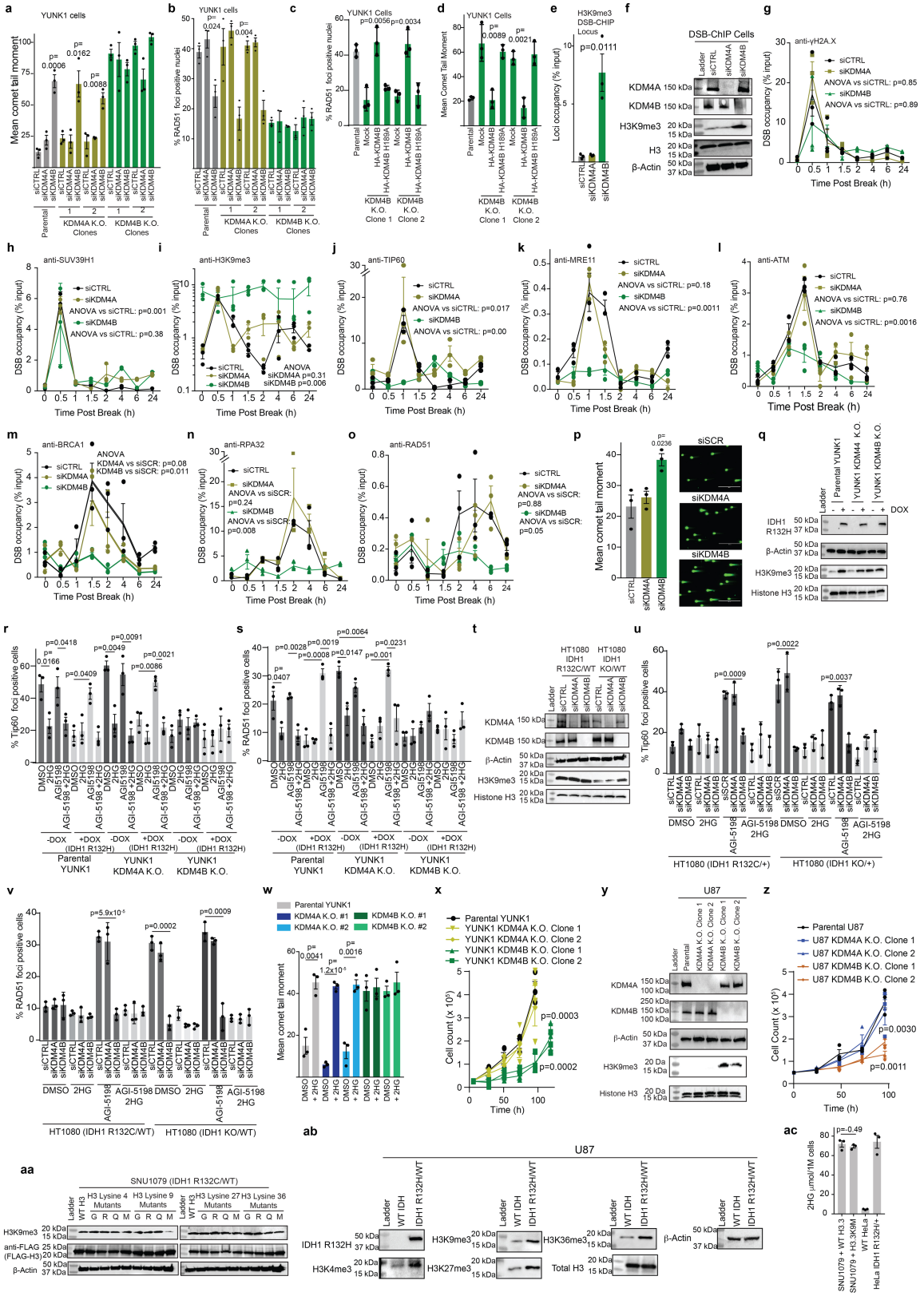
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | αKG supplementation rescues oncometabolite-induced HDR deficiency. a, b, Western blot analysis of H3K9me3 levels in the following cells with or without treatment with 2 mM αKG for 96 h: SNU1079 *IDH1*$^{R132C/+}$ cholangiocarcinoma cells (**a**), U87 *IDH1*^{WT} and *IDH1*$^{R132H/+}$ glioblastoma cells (**b**), UOK 262 *FH*$^{-/-}$ renal cell carcinoma cells with and with *FH* cDNA complementation, and YUNK1 cells with shRNA suppression of *SDHB* (shSDHB) or *FH* (shFH). These experiments were repeated twice with similar results. **c**, Quantification of cells with TIP60 foci-positive nuclei (>10 foci per nucleus) at 1 h post 2 Gy ionizing radiation in the following cells with or without treatment with 2mM αKG for 96 h: *IDH1*^{WT} and *IDH1*$^{R132H/+}$ U87 glioblastoma cells, *IDH1*^{WT} and *IDH1*$^{R132H/+}$ HeLa cells, UOK 262 *FH*$^{-/-}$ renal cell carcinoma cells, YUNK1 cells with shRNA suppression of *SDHB* (shSDHB) or *FH* (shFH), compared to non-targeting control shRNA (shCTRL), and HEK293FT cells with shRNA suppression of *SDHB* (shSDHB) or *FH* (shFH) compared to non-targeting control shRNA (shCTRL). **d**–**f**, Quantification of cells with pATM S1918 foci-positive nuclei (>10 foci per nucleus) at 1 h after 2 Gy ionizing radiation in the following cells treated or not with 2 mM αKG for 48 h as indicated: HEK293FT cells with shRNA suppression of *SDHB* (shSDHB) or *FH* (shFH) compared to non-targeting control shRNA (shCTRL) (**d**), *IDH1*^{WT} and *IDH1*$^{R132H/+}$ HeLa cells (**e**), and YUNK1 cells with shRNA suppression of *SDHB* (shSDHB) or *FH* (shFH) compared to non-targeting control shRNA (shCTRL) (**f**). **g**, Quantification of cells with RAD51 foci-positive nuclei (>10 foci per nucleus) at 4 h post 2 Gy ionizing radiation in U87 *IDH1*^{WT} and *IDH1*$^{R132H/+}$ glioblastoma cells treated with either 2 mM αKG for 48 h or with DMSO control. **h**, Quantification of cells with RAD51 foci-positive nuclei (>10 foci per nucleus) at 4 h post 2 Gy ionizing radiation in SNU1079 (*IDH1*$^{R132C/+}$) cholangiocarcinoma cells pre-treated with DMSO, 2 mM αKG or AGI-5198 for 48 h. **i**, Western blot analysis of H3K9me3 and total H3 levels in U2OS EJ-DR cells treated with DMSO (control), 2 mM αKG, 500 μM octyl-R-2HG, 2 mM succinate, 30 μM dimethyl fumarate, or the indicated combinations of αKG plus 2HG, succinate or fumarate. This experiment was repeated twice with similar results. **j**, ChIP analysis of H3K9me3 occupancy at the DSB–ChIP reporter locus in U2OS cells in the absence of a DSB after treatment with DMSO (control), 500 μM octyl-R-2HG, 2 mM succinate or 30 μM dimethyl fumarate, in all cases with or without 2 mM αKG, as indicated. **k**–**r**, Heat maps of the relative occupancy of the indicated factors at the site-directed DSB in U2OS cells as measured by ChIP and normalized to the uninduced controls. The assay was performed at the indicated time points post addition of Shield-1 and triamcinolone in DMSO-treated cells (control) (**k**), cells treated with 2 mM αKG (**l**), cells treated with 500 μM octyl-(R)-2HG (**m**), cells treated with 2 mM αKG and 500 μM octyl-(R)-2HG (**n**), cells treated with 2 mM succinate (**o**), cells treated with 2 mM succinate and 2 mM αKG (**p**), cells treated with 30 μM dimethyl fumarate (**q**), and cells treated with 30 μM dimethyl fumarate and 2 mM αKG (**r**). The heat maps for 2HG alone, succinate alone, fumarate alone and DMSO control alone are reproduced from Fig. 2b and are presented again here for comparison. **s**–**aa**, Line graphs of percent input values for DSB–ChIP assays with antibodies (corresponding to the heat maps in **l**, **n**, **p**, **r**) for γH2A.X (**s**; αKG + 2HG, $F = 0.04$, df = 1; αKG + succinate, $F = 2.76$, df = 1; αKG + fumarate, $F = 0.18$, df = 1); SUV39H1 (**t**; αKG + 2HG, $F = 0.73$; df = 1, αKG + succinate, $F = 0.55$, df = 1; αKG + fumarate, $F = 0.09$, df=1); H3K9me3 (**u**; αKG + 2HG, $F = 0.076$, df = 1; αKG + succinate, $F = 4.05$, df = 1; αKG + fumarate, $F = 8.910$, df = 1); TIP60 (**v**; αKG + 2HG, $F = 1.32$, df = 1; αKG + succinate, $F = 1.98$, df = 1; αKG + fumarate, $F = 107.8$, df = 1); MRE11 (**w**; αKG + 2HG, $F = 0.53$, df = 1; αKG + succinate, $F = 1.2$, df = 1; αKG + fumarate, $F = 2.3$, df = 1); ATM (**x**; αKG + 2HG, $F = 14.8$, df = 1, αKG + succinate, $F = 0.31$, df = 1; αKG + fumarate, $F = 8.67$, df = 1); BRCA1 (**y**; αKG + 2HG, $F = 3.3$, df = 1; αKG + succinate, $F = 2.1$, df = 1; αKG + fumarate $F = 1.5$, df = 1); RPA32 (**z**; αKG + 2HG, $F = 0.003$, df = 1; αKG + succinate, $F = 1.78$, df = 1, αKG + fumarate, $F = 0.57$, df = 1); and RAD51 (**aa**; αKG + 2HG, $F = 1.4$, df = 1; αKG + succinate, $F = 2.4$, df = 1; αKG + fumarate, $F = 3.10$, df = 1) at the indicated time points after addition of triamcinolone and Shield-1 to induce an I-SceI break in the U2OS DSB–ChIP cells. Line graphs corresponding with **k**, **m**, **o**, **q**, are presented in Extended Data Fig. 3b–j. **ab**, Quantification of HDR efficiency as measured by restoration of a functional GFP gene following I-SceI induction of a DSB in the DR–GFP reporter in U2OS cells following pre-treatment with DMSO, 500 μM octyl-R-2HG, 2 mM succinate or 30 μM dimethyl fumarate, and with no αKG, 1 mM αKG or 2 mM αKG as indicated. In **c**–**h**, **j**, and **ab**, data are mean ± s.e.m. with $n = 3$ biological replicates; statistical analysis by two-tailed unpaired *t*-test; df=4. In **s**–**aa**, lines run through the mean ± s.e.m. with $n = 3$ biological replicates for each time point; statistical analysis by ANOVA; *P* values are indicated.



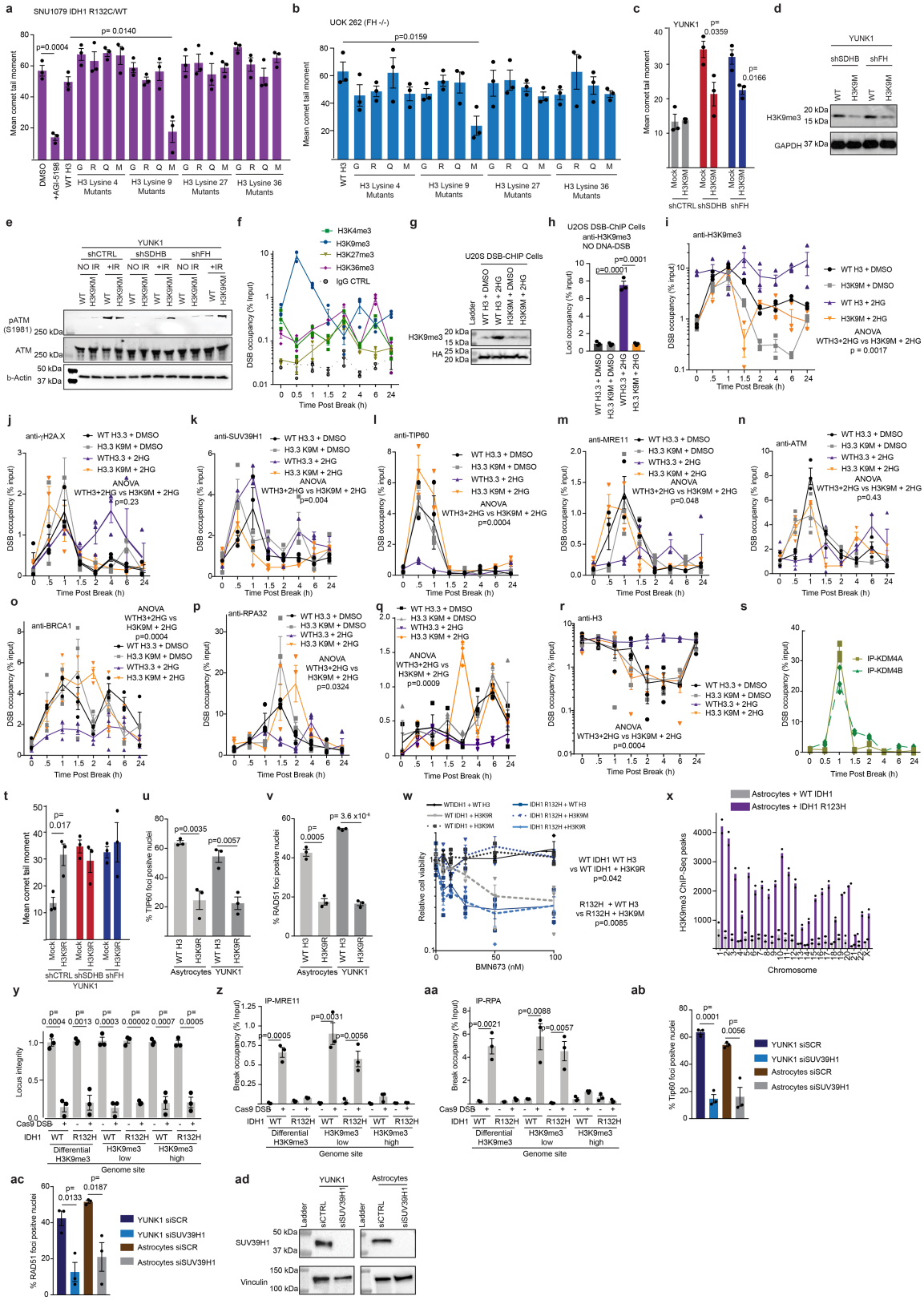
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Overexpression of KDM4A and KDM4B suppresses elevated H3K9me3 and rescues HDR deficiency in oncometabolite-producing cells. a–**c**, Western blot analysis of H3K9me3 and total H3 levels in the cholangiocarcinoma cell line SNU1079 (*IDH1*$^{R132C/+}$) (**a**), and renal cell carcinoma cell lines UOK262 (*FH*$^{-/-}$) (**b**) and NCCFH1 (*FH*$^{-/-}$) (**c**) treated as indicated or after transfection with vectors for expression of haemagglutinin (HA)-tagged KDM4A, catalytically inactive KDM4A(H188A), KDM4B, catalytically inactive KDM4B(H189A), KDM4C and KDM6A or vectors for expression of Flag-tagged ALKBH2, ALKBH3 and JMJD4. This experiment was repeated twice with similar results. **d**, Quantification of RAD51-positive foci in SNU1079 cells after the indicated treatment or transfection. **e**–**j**, Quantification of TIP60 foci-positive nuclei (>10 foci per nucleus) (**e**, **f**), RAD51 foci-positive nuclei (**g**, **h**), and mean comet-tail moment in UOK 262 and NCCFH1 *FH*$^{-/-}$ renal cell carcinoma cell lines (**i**, **j**), after transfection with FH expression constructs or after with vectors for expression of HA-tagged KDM4A, catalytically inactive KDM4A(H188A), KDM4B, catalytically inactive KDM4B(H189A), KDM4C, or KDM6A or vectors for expression of Flag-tagged ALKBH2, ALKBH3 or JMJD4. In **g**, **h**, cells treated with 2 mM αKG or DMSO control are also included in the analysis. **k**, Western blots showing H3K9me3 and total H3 levels after transfection with constructs for expression of HA-tagged KDM4A or KDM4B in YUNK1 shSDHB or shFH cells. This experiment was repeated twice with similar results. **l**, Quantification of neutral comet assay in immortalized astrocytes expressing IDH1 or IDH1(R132H) and in YUNK1 shCTRL, shSDHB and shFH cells, with or without overexpression of KDM4A or KDM4B as indicated, or with 24 h treatment with 2 mM αKG. **m**, **n**, Quantification of RAD51 nuclear foci 6 h after 2 Gy ionizing radiation in *IDH1*^{WT} or *IDH1*$^{R132/+}$ HeLa cells (**m**) and in HEK293FT cells (**n**) with shRNA suppression of *SDHB* (shSDHB) or *FH* (shFH), compared with non-targeting shRNA (shCTRL). Cells were irradiated after 24 h pretreatment with αKG or 24 h after transfection with expression vectors for KDM4A or KDM4B. **o**, **p**, Quantification of BRCA1 nuclear foci 4 h after 2 Gy ionizing radiation in *IDH1*^{WT} or *IDH1*$^{R132H/+}$ HeLa cells (**o**) and HEK293FT cells (**p**) with shRNA suppression of *SDHB* (shSDHB) or *FH* (shFH), compared with non-targeting shRNA (shCTRL). Cells were irradiated after 24 h pretreatment with αKG or 24 h after transfection with expression vectors for KDM4A or KDM4B. **q**, Quantification of neutral comet assays performed in WT and *HIF1A*-knockout mouse embryonic fibroblasts (MEFs) after treatment with 500 μM octyl-(R)-2HG, 2 mM succinate or 30 μM dimethyl fumarate, compared to DMSO control. **r**, Western blot analysis of HIF-1α after indicated treatment of WT and *HIF1A*-knockout MEFs with 500 μM octyl-(R)-2HG, 2 mM succinate or 30 μM dimethyl fumarate. Hypoxia exposure at 1% O_2 for 24 h is used as a positive control for HIF-1α stabilization. **s**, Western blot analysis of KDM4A and KDM4B expression and H3K9me3 levels in *KDM4A*-knockout and *KDM4B*-knockout YUNK1 cell lines. This experiment was repeated twice with similar results. **t**, **u**, Quantification of RAD51 foci-positive cells 4 h after 2 Gy ionizing radiation (**t**) and comet-tail moment (**u**) in parental YUNK1 cells, *KDM4A*-knockout YUNK1 cells, and *KDM4B*-knockout YUNK1 cells with overexpression constructs for KDM4A or KDM4B or mock transfection, as indicated. **v**, Western blot analysis of KDM4A and KDM4B expression and of H3K9me3 levels in *KDM4A*- and *KDM4B*-knockout YUNK1 cells compared to parental YUNK1 controls after transfection with overexpression constructs for HA-tagged *KDM4A* and *KDM4B* open reading frames, as indicated. This experiment was repeated twice with similar results. In **d**–**h**, **j**, **l**–**q**, **t**, **u**, data are mean ± s.e.m. with $n = 3$ biological replicates; statistical analysis by two-tailed unpaired *t*-test; df = 4, *P* values are indicated.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Inhibition of KDM4B mediates oncometabolite-induced HDR deficiency. a, Quantification of neutral comet assay (**a**) and quantification of RAD51 foci-positive cells (>10 foci per nucleus) (**b**) 4 h after 2 Gy ionizing radiation in parental YUNK1 cells, *KDM4A*-knockout YUNK1 cells, and *KDM4B*-knockout YUNK1 cells after siRNA suppression of either *KDM4A* or *KDM4B*, or non-targeting siRNA control (siCTRL), as indicated. **c**, Quantification of RAD51 foci-positive cells (**c**) and quantification of neutral comet assay (**d**) in *KDM4B*-knockout YUNK1 cells transfected with expression constructs for either KDM4B(WT) or the catalytically inactive KDM4B(H189A). **e**, Quantification by ChIP of baseline H3K9me3 levels (in the absence of a DSB) at the DSB–ChIP reporter locus in U2OS cells after siRNA suppression of either *KDM4A* or *KDM4B*, compared to non-targeting control siRNA. **f**, Validation of siRNA suppression of *KDM4A* and *KDM4B* and documentation of H3K9me3 levels by western blot in the U2OS DSB–ChIP cells. This experiment was repeated twice with similar results. **g**–**o**, Per cent input values for DSB–ChIP assays performed after siRNA suppression of KDM4A or KDM4B with antibodies (corresponding to Fig. 3g–i) for γH2A.X (**g**; siKDM4A, $F = 0.0$, df = 1; siKDM4B, $F = 0.02$, df = 1); SUV39H1 (**h**; siKDM4A, $F = 60.85$, df = 1; siKDM4B, $F = 0.98$, df = 1); H3K9me3 (**i**; siKDM4A, $F = 1.4$, df = 1; siKDM4B, $F = 28.3$, df = 1); TIP60 (**j**; siKDM4A, $F = 15.2$, df = 1; siKDM4B, $F = 41.3$, df = 1); MRE11 (**k**; siKDM4A, $F = 15.5$, df = 1; siKDM4B, $F = 69.3$, df = 1); ATM (**l**; siKDM4A, $F = 0.1$, df = 1; siKDM4B, $F = 15.4$, df = 1); BRCA1 (**m**; siKDM4A, $F = 5.5$, df = 1; siKDM4B, $F = 19.94$, df = 1); RPA32 (**n**; siKDM4A, $F = 1.9$, df = 1; siKDM4B, $F = 24.5$, df = 1); and RAD51(**o**; siKDM4A, $F = 0.88$, df = 1; siKDM4B, $F = 7.4$, df = 1) at the indicated time points after addition of triamcinolone and Shield-1 to induce an I-SceI break in DSB–ChIP U2OS cells. **p**, Quantification and representative images of neutral comet assays performed in DSB–ChIP U2OS cells after siRNA suppression of *KDM4* or *KDM4B* compared to a non-targeting control siRNA (siCTRL). Scale bars, 400 μm. **q**, Western blot analysis of IDH1(R132H) expression in parental YUNK1, *KDM4A*-knockout YUNK1 cells and *KDM4B*-knockout YUNK1 cells treated with either doxycycline (DOX; to induce expression of IDH1(R132H)) or vehicle control, and western blot analysis of global H3K9me3 and total H3 levels. This experiment was repeated twice with similar results. **r**, **s**, Quantification of TIP60 (**r**) and RAD51 (**s**) foci-positive cells (>10 foci per nucleus) after 2 Gy ionizing radiation (at 1 h for TIP60 and 4 h for RAD51) in parental, *KDM4A*-knockout and *KDM4B*-knockout YUNK1 cells, treated with either doxycycline or vehicle control, and also treated as indicated with DMSO, 500 μM octyl-(R)-2HG, 1 mM AGI-5198 or both 500 μM octyl-(R)-2HG and 1 mM AGI-5198. **t**, Western blot analysis of KDM4A and KDM4B levels in HT1080 cells ($IDH1^{R132C/+}$) and in HT1080 cells with CRISPR–Cas9 knockout of the mutant IDH1 allele ($IDH1^{KO/+}$) transfected with scramble siRNA control (siSCR), siKDM4A or siKDM4B, and western blot analysis of global H3K9me3 and total H3 levels. This experiment was repeated twice with similar results. **u**, **v**, Quantification of TIP60 (**u**) and RAD51 (**v**) foci-positive cells (>10 foci per nucleus) after 2 Gy ionizing radiation in HT1080 cells ($IDH1^{R132C/+}$) and in HT1080 cells with CRISPR–Cas9 knockout of the mutant $IDH1^{R132C}$ allele ($IDH1^{KO/+}$) transfected with siSCR, siKDM4A or siKDM4B, and treated with or without 500 μM 2HG, 1 mM AGI-5198 or both 1 mM AGI-5198 and 500 μM 2HG. **w**, Quantification of mean comet-tail moment in parental, *KDM4A*-knockout and *KDM4B*-knockout YUNK1 cells treated with either 500 μM octyl-R-2HG or DMSO control. **x**, Quantification of cell proliferation by serial cell counts over time of parental, *KDM4A*-knockout (two independent clones) and *KDM4B*-knockout YUNK1 (two independent clones) cells. **y**, Western blot analysis of KDM4A and KDM4B expression levels in U87 glioma cells with *KDM4A* or *KDM4B* knockout, compared to parental U87 cells, and western blot analysis of global H3K9me3 and total H3 levels. This experiment was repeated twice with similar results. **z**, Quantification of cell proliferation by serial cell counts of parental, *KDM4A*-knockout (two independent clones) and *KDM4B*-knockout (two independent clones) U87 cells. **aa**, Western blot analysis of expression of mutant H3 constructs and analysis of H3K9me3 levels after expression of the indicated H3 mutants in SNU1079 $IDH1^{R132C}$ cholangiocarcinoma cells. Note that only H3K9M can reduce global H3K9me3 levels in *IDH1*-mutant cells. This experiment was repeated twice with similar results. **ab**, Western blot analysis of IDH1(R132H), H3K4me3, H3K9me3, H3K27me3, H3K36me3 and total H3 in IDH^{WT} and $IDH1^{R132H/+}$ U87 glioblastoma cells. This experiment was repeated twice with similar results. **ac**, Quantification of 2HG levels by fluorometric 2HG detection assay in SNU1079 ($IDH1^{R132C/+}$) cholangiocarcinoma cells transfected with either H3.3(WT) or H3.3(K9M) expression constructs. In **a**–**e, p**, **r**, **s**, **u**–**w**, **ac**, data are mean ± s.e.m. with $n = 3$ biological replicates; statistical analysis by two-tailed unpaired *t*-test; df = 4. In **f**, **h**, **i**–**q**, data are mean ± s.e.m. with $n = 3$ biological replicates; statistical analysis by ANOVA; *P* values as indicated.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Oncometabolites induce HDR deficiency via hypermethylation of H3K9. a, **b**, Quantification of neutral comet assay performed in SNU1079 (*IDH1*$^{R132C/+}$) cholangiocarcinoma cells (**a**) and UOK262 *FH*$^{-/-}$ renal cell carcinoma cells (**b**) after expression of the indicated H3.3 mutants. **c**, Quantification of neutral comet assay performed in YUNK1 shSDHB, shFH or shCTRL cells transfected with expression constructs for H3.3(WT) versus H3.3(K9M). Cells were assayed 24 h after transfection with the indicated construct. **d**, Western blot analysis of global H3K9me3 levels in YUNK1 shSDHB or shFH cells transfected with vector for expression of H3.3(WT) or H3.3(K9M). This experiment was repeated twice with similar results. **e**, Western blot analysis of pATM and total ATM levels in shCTRL, shSDHB and shFH YUNK1 cells transfected with expression constructs for H3.3(WT) or H3.3(K9M), at 1 h after ionizing radiation (5 Gy) or without ionizing radiation. This experiment was repeated twice with similar results. **f**, Quantification over time by ChIP of H3K9me3, H3K4me3, H3K27me3 and H3K36me3 levels after induction of the site-specific DSB in U2OS DSB–ChIP cells after DMSO treatment. The H3K9me3 data is also presented in Fig. 2 and Extended Data Fig. 3 for relevant comparisons. **g**, Western blot analysis of H3K9M mutant construct expression and H3K9me3 levels in U2OS DSB–ChIP cells after expression of H3K9M or WT H3, or treatment with DMSO control or 500 μM octyl-(R)-2HG. This experiment was repeated twice with similar results. **h**, ChIP analysis of H3K9me3 occupancy at the DSB–ChIP locus in U2OS cells (in the absence of an induced DSB), after transfection with construct for expression of H3.3(WT) or H3.3(K9M) expression construct after indicated treatment with either 500 μM 2HG or DMSO control. **i**–**q**, Per cent input values for assays performed in U2OS cells after transfection with expression construct for H3.3(WT) or H3.3(K9M) and treated with 500 μM 2HG or DMSO as indicated (corresponding to Fig. 4b–e), with antibodies for H3K9me3 (**i**; $F = 54.63$, df = 1), γH2A.X (**j**; $F = 1.95$, df = 1), SUV39H1 (**k**; $F = 34.11$, df = 1), TIP60 ($F = 126.6$, df = 1), MRE11 (**m**; $F = 7.9$, df=1), ATM (**n**; $F = 0.75$, df = 1), BRCA1 (**o**; $F = 119.5$, df = 1), RPA32 (**p**; $F = 10.34$, df = 1), RAD51 (**q**; $F = 80.2$, df = 1) and total H3 (r; $F = 120.6$, df = 1) at the indicated time points after addition of triamcinolone and Shield-1 to induce an I-SceI break. **s**, Line graphs of percent input values for DSB–ChIP assays performed with antibodies for KDM4A and KDM4B in U2OS cells at the indicated time points after addition of triamcinolone and Shield-1 to induce an I-SceI break. **t**, Quantification of neutral comet assay performed in YUNK1 cells with shSDHB, shFH or shCTRL, transfected with constructs for expression of H3.3(WT) or H3.3(K9R). Cells were assayed 24 h after transfection with the indicated construct. **u**, **v**, Quantification of TIP60 foci-positive nuclei (>10 foci per nucleus) (**u**) at 1 h post 2 Gy ionizing radiation and RAD51 foci-positive nuclei at 4 h post 2 Gy ionizing radiation (**v**) in immortalized astrocytes and YUNK1 cells transfected with constructs for expression of H3.3(WT) or H3.3(K9R). **w**, PARP inhibitor sensitivity in *IDH1*^{WT} and *IDH1*$^{R132H/+}$ U87 glioblastoma cells transfected with expression constructs for H3.3(WT), H3.3(K9M) or H3.3(K9R). (*IDH*^{WT}: H3.3(WT) vs H3.3(K9R), $F = 8.7$, df = 1; *IDH1*^{R132H}: H3.3(WT) vs H3.3(K9M), $F = 23.0$, df = 1). **x**, Quantification of genomic H3K9me3 peaks across chromosomes by analysis of ChIP-sequencing data for H3K9me3 in a matched pair of immortalized human astrocyte cell lines expressing IDH1(WT) or IDH1(R132H)[28]. Peaks were called using HOMER (v.4.10) and are defined as spanning at least 1,000 base pairs and at least 2,500 base pairs apart, filtered by $P \leq 0.01$. $n = 2$ technical replicates. **y**, qPCR analysis of amplicons that span the Cas9–guide RNA cleavage target sites to asses I-SceI cleavage at the H3K9me3 differentially methylated, H3K9me3 both low, and H3K9me3 both high loci in *IDH1*^{WT} and *IDH1*$^{R132H/+}$ astrocytes. Reduced amplification of the genomic DNA across the Cas9–guide RNA target sites indicates that cleavage at the target site has occurred. **z**, **aa**, ChIP analysis of MRE11 (**z**) and RPA32 (**aa**) at H3K9me3 differentially methylated, H3K9me3 low, and H3K9me3 high loci as identified in **x**, in the *IDH1*^{WT} and *IDH1*^{R132H} astrocyte cell lines 12 h after Cas9 nucleofection. **ab**–**ac**, Quantification of TIP60 foci-positive nuclei (>10 foci per nucleus) at 1 h after 2 Gy ionizing radiation (**ab**) and RAD51 foci-positive nuclei at 4 h after 2 Gy ionizing radiation (**ac**) in immortalized astrocytes and YUNK1 cells transfected with siRNA to knock down *SUV39H1* or non-targeting control (siCTRL). **ad**, Western blot analysis of SUV39H1 levels in YUNK1 cells and immortalized astrocytes transfected with siRNA to knock down *SUV39H1* or non-targeting control siRNA (siCTRL). In **f**, **i**–**s**, **w**, data are mean ± s.e.m.; statistical analysis by ANOVA, with *P* values as indicated. In **a**–**c**, **h**, **t**–**v**, **y**–**ac**, data are mean ± s.e.m. with $n = 3$ biological replicates; statistical analysis by two-tailed unpaired *t*-test; df = 4, *P* values are indicated.

nature research

Corresponding author(s): Peter Glazer

Last updated by author(s): Mar 7, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All software is commercially or freely available. Focinator V 2.0 was used to analyze Immunfluorescence images. Statistical Analysis was done in PRISM 7 or Stata. Figures were compiled using Adobe Illustrator. For LC/MS assays, Analyst® 1.6 software was used for system control and data acquisition and MultiQuant® 3.0 software was used for data processing and quantitation. Flow cytometry data was collected using FlowJo CE and analyzed using FlowJo™ v7.6.5 software.
Data analysis	All software is commercially or freely available. Focinator V 2.0 was used to analyze Immunfluorescence images. Statistical Analysis was done in PRISM 7 or Stata. Figures were compiled using Adobe Illustrator. Analyst® 1.6 software was used for system control and data acquisition and MultiQuant® 3.0 software was used for data processing and quantitation. Flow cytometry data was analyzed using FlowJo™ v7.6.5 software. For ChIP-seq analysis, raw sequencing data was aligned to the GRhg38 (hg38) genome build using Bowtie2 (v2.3.4.1). H3K9me3 peaks were called using HOMER (v4.10) using findPeaks with options -style histone -size 1000 -minDist 2500, filtered by P = 0.01.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated or analyzed during this study are included in this published article (or its supplementary information files, including source data). There are no restrictions on data availability.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For in vitro assays, n=3 biological replicates were used for sample size in all cases order to allow statistical analyses. Each biological replicate was defined as an independent culture of cells. No statistical methods were used to determine sample size. For the in vivo mouse studies, mice were randomized into irradiated and mock irradiated groups with 8 mice/group, based on a calculated power of 80% to detect a difference of 1.25 standard deviations between two groups at 5% significance level.
Data exclusions	No data were excluded.
Replication	Replicates are indicated in the manuscript with n=3 biological replicates for experimental data presented here. All replicates were successful.
Randomization	Cultures and mice were randomly assigned to either experimental or control groups.
Blinding	In vitro assays were not conducted in a blinded manner. Xenograft tumors were randomly assigned into treatment or control groups. Tumor growth measurements were not conducted in a blinded fashion. Blinding was not relevant to our study, since subjective rating of data was not involved.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

anti-ATM (phospho S1981) (abcam, ab36810)
anti-ATM(Millipore, Ab-3)
anti ß-Actin HRP (Proteintech, HRP-60008)
anti-Histone 3 Lysine 9 trimethyl (D4W1U, Cell Signaling Technology)
anti-Histone H3 (# 9715 Cell Signaling)
anti-Histone H3 (ab1791, Abcam)
Rabbit anti-HA-Tag (C29F4, Cell Signaling Technology)
mouse anti- GAPDH HRP (HRP-60004, Proteintech),
mouse anti-IDH1 R132H (H09, Dianova)
anti-IDH1 (D2H1, Cell Signaling Technology)
anti-Fumarase (D9C5, Cell Signaling Technologies)
anti-SDHB (21A11AE7, Abcam)
anti-RAD51(14B4 Novus bio)
anti-BRCA2 (ab1, Millipore)
ati-TI60 (NBP2-24613. Novus)
anti-RPA32 (#52448, Cell Signaling Technology)
anti-MRE11 (H-300, Sant Cruz Biotechnology sc-22767)
anti-Phospho-(Ser/Thr) ATM/ATR Substrate Antibody (Cell Signaling Technology, #2851)
anti-KDM4A (A300-861A ,Bethyl)
anti-KDM4A (C70G6, Cell Signaling Technology)
anti KDM4B (A301-478A, Bethyl) 1:1000 5% BSA

anti-KDM4B (D7E6, Cell Signaling Technology)
anti-Ku80 80 (Clone 7, BD Biosciences)
anti-Poly-ADP-Ribose (PAR) (4335-mc-100, Trevigen)
anti-phospho-RPA32 (E5A2F, Cell Signaling Technology)
Mouse anti-CHK1 (2G1D5, Cell Signaling Technology)
anti-phospho-CHK1 (#2341 Cell Signaling Technology)
anti-phospho-CHK2 (C13C1, Cell Signaling Technologies)
anti-HIF1a (NB100-105, Novus)
anti-Vinculin (ab129002, abcam)
anti Histone H2A.X (D17A3, Cell Signaling Technology)
anti DNA-PKcs (ab70250, abcam)
anti phosphor-DNA-PKcs (ab 18192, abcam)
anti-H3K36me3 (D5A7, Cell Signaling Technologies)
anti-H3K27me3 (C36B11, Cell Signaling Technologies)
Rabbit anti-H3K4me3 (C42D8, Cell Signaling Technologies)

Antibody dilutions for western blots were as follows: mouse anti-ATM (phospho S1981) antibody (Abcam, ab36810) 1:100 in 5% milk, rabbit anti-ATM (Millipore, Ab-3) 1:1000 in 5% milk, mouse anti-ß-Actin HRP (Proteintech, HRP-60008) 1:1000 in 5% milk, anti-Histone 3 Lysine 9 trimethyl (D4W1U, Cell Signaling Technology) 1:1000 in 5% BSA, rabbit polyclonal anti-Histone H3 (ab1791, Abcam) 1:5000 in 5% BSA, rabbit anti-HA-Tag (C29F4, Cell Signaling Technology) 1:2000 in 5% BSA, mouse anti-GAPDH HRP (HRP-60004, Proteintech), mouse anti-IDH1 R132H (H09, Dianova) 1:2000 in 5% milk, rabbit anti-IDH1 (D2H1, Cell Signaling Technology) 1:1000 in 5% BSA, rabbit monoclonal anti-Fumarase (D9C5, Cell Signaling Technologies) 1:1000 in 5% BSA, mouse monoclonal anti-SDHB (21A11AE7, Abcam) 1:1000 in 5% milk, mouse monoclonal anti-RAD51(14B4 Novus bio) 1:1000 in 5% milk, mouse anti-BRCA2 (ab1, Millipore) 1:1000 in 5% milk, rabbit anti-T60 (NBP2-24613, Novus Biologicals) 1:1000 in 5% BSA, rabbit anti-RPA32 (#52448, Cell Signaling Technology) 1:1000 in 5% BSA, Rabbit anti -MRE11 (H-300, Sant Cruz Biotechnology sc-22767) 1:1000 in 5% milk, rabbit anti- Phospho-(Ser/Thr) ATM/ATR Substrate Antibody (#2851,Cell Signaling Technology) 1:1000 in 5% BSA, rabbit anti-KDM4A (A300-861A, Bethyl) 1:1000 in 5% BSA, rabbit anti-KDM4A (C37E5, Cell Signaling Technology) 1:1000 in 5% BSA, anti-KDM4B (A301-478A, Bethyl) 1:1000 in 5% BSA, anti-KDM4B (D7E6, Cell Signaling Technology) 1:1000 in 5% BSA, rabbit anti-FLAG (SAB4301135, Sigma) 1:1000 in 5% BSA, rabbit anti-SUV39H1 (D11B6, Cell Signaling Technology) 1:1000 in 5% BSA, mouse anti-Ku80 80 (Clone 7, BD Biosciences) 1:1000 in 5% Milk, rabbit anti-Histone H2A.X (D17A3, Cell Signaling Technology) 1:1000 in 5% BSA, mouse anti-Poly-ADP-Ribose (PAR) (4335-mc-100, Trevigen), mouse anti-phospho-ATM S1981 (ab36810, Abcam) 1:1000 in 5% Milk, rabbit anti-phospho-RPA32 (E5A2F, Cell Signaling Technology) 1:1000 in 5% BSA, mouse anti-CHK1 (2G1D5, Cell Signaling Technology) 1:1000 in 5% BSA, rabbit anti-phospho-CHK1 (#2341, Cell Signaling Technology) 1:1000 in 5% BSA, rabbit anti-CHK2 (#2662, Cell Signaling Technologies) 1:1000 in 5% BSA, rabbit anti-phospho-CHK2 (C13C1, Cell Signaling Technologies) 1:1000 in 5% BSA, rabbit anti-DNA-PKcs (ab70250, Abcam) 1:1000 in 5% milk, rabbit anti phospho-DNA-PKcs (ab 18192, Abcam) 1:1000 in 5% milk, mouse anti-HIF1a (NB100-105, Novus) 1:1000 in 5% BSA, rabbit anti-H3K36me3 (D5A7, Cell Signaling Technologies) 1:1000 in 5% BSA, rabbit anti-H3K27me3 (C36B11, Cell Signaling Technologies) 1:1000 in 5% BSA, rabbit anti-H3K4me3 (C42D8, Cell Signaling Technologies) 1:1000 in 5% BSA.

Validation

Antibody validation appears in Extended Data Figure 3. We used siRNA to knockdown the target proteins and validated knockdown by western blot, then we performed the DSB-ChIP protocol in the cells with the target protein suppressed to validated the ChIP signal from these antibodies. Knockdown of the H3K9 specific methyltransferase SUV39H1 served to validate the H3K9me3 antibody. All antibodies used in this study are commercially available, and were otherwise validated by the manufacturer, by previous studies from other laboratories or by previous studies from our laboratory, as cited in the text and methods.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

YUNK1 cells were derived by us and have been previously described (6). They were generated in culture from uninvolved cortical renal tissue from a patient undergoing a radical nephrectomy for kidney cancer. They were immortalized with a lentiviral vector containing SV40 Large T antigen (Addgene plasmid #22298). YUNK1 cells with shRNA suppression of SDHB or FH using the GIPZ lentiviral constructs (GE Dharmacon) have been previously described (6). shRNA suppression of SDHB and FH in the HEK293FT cells (obtained from ATCC) were achieved using TRIPZ lentiviral knockdowns and these cells are previously described (6). Hela cells were obtained from ATCC. Heterozygous IDH1 R132H/+ HeLa cells were derived by us and have been previously described (5). HCT116 cells and HCT116 IDH1 R132H/+ cells were obtained from Horizon Discovery (Waterbeach, UK). Immortalized astrocytes expressing WT or mutant IDH1 have been previously described (9,28) and were a gift from T. Chan (Sloan Kettering). U87 IDH1 WT and IDH1 R132H/WT glioblastoma cells were obtained from the ATCC. SN1079 and RBE cells were obtained from the RIKEN cell bank. SW1353 cells were obtained from ATCC. UOK 262 cells were a gift from M. Linehan and have been previously described (29). NCCFH1 were obtained from B.T. Teh. and have been previously described (30).

Authentication

Cell lines were authenticated using STR analysis.

Mycoplasma contamination

Cell were routinely tested for mycoplasma and tested negative.

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified cell lines were used in this study.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Female athymic nu/nu mice (Hsd:Athymic Nude-Foxn1nu, Envigo) were used at 8 weeks of age.
Wild animals	not applicable
Field-collected samples	not applicable
Ethics oversight	The authors confirm that all animal experiments were performed in accordance with relevant guidelines and regulations. All studies were approved by the Yale University Institutional Animal Care and Use Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	We did not generate ChIP-seq data in this study. Rather, we analyzed published and publicly available data. Raw ChIP-seq data corresponding to dox-induced IDH1 R132H mutant (MUT) and uninduced parental (PAR) astrocyte cell lines after 40 passages in culture, as well as corresponding input data, were accessed from publicly available SRA data (SRP082568) from experiments previously published in Turcan, S. et al. Mutant-IDH1-dependent chromatin state reprogramming, reversibility, and persistence. Nat Genet 50, 62-72, doi:10.1038/s41588-017-0001-z (2018).
Files in database submission	Not applicable. See comment above.
Genome browser session (e.g. UCSC)	not applicable

Methodology

Replicates	Not applicable. See comment above.
Sequencing depth	Not applicable. See comment above.
Antibodies	Not applicable. See comment above.
Peak calling parameters	Not applicable. See comment above.
Data quality	Not applicable. See comment above.
Software	See comment above. For ChIP-seq analysis, data from Turcan, S. et al. Mutant-IDH1-dependent chromatin state reprogramming, reversibility, and persistence. Nat Genet 50, 62-72, doi:10.1038/s41588-017-0001-z (2018) was aligned to the GRhg38 (hg38) genome build using Bowtie2 (v2.3.4.1). H3K9me3 peaks were called using HOMER (v4.10) using findPeaks with options -style histone -size 1000 -minDist 2500, filtered by P = 0.01.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	For the EJ2 and EJ5 assays, reporter cells were washed with PBS, collected by trypsinization, then assayed for GFP expression by flow cytometry. For cell cycle analysis, cells were fixed with cold ethanol for 1 h at 4C, then stained with Propidium iodide solution.
Instrument	Becton Dickinson FACSCalibur

Software	Data was collected with FlowJo CE and data was analyzed with FlowJo version 7.6.5 software.
Cell population abundance	Cell populations for EJ2 and EJ5 were differentiable by GFP expression above 10^3 intensity and quantified with FlowJo.
Gating strategy	For GFP analysis in the EJ2 and EJ5 assay, the GFP threshold for gating was established using the positive control samples and set at 10^3 intensity. This gating strategy is shown in the supplementary Information. A negative control sample confirmed this gating strategy, before the exact same gate was applied to all experimental samples.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

The proteome landscape of the kingdoms of life

https://doi.org/10.1038/s41586-020-2402-x

Received: 2 August 2019

Accepted: 27 April 2020

Published online: 17 June 2020

Check for updates

Johannes B. Müller[1,7], Philipp E. Geyer[1,2,7], Ana R. Colaço[3], Peter V. Treit[1], Maximilian T. Strauss[1,2], Mario Oroshi[1], Sophia Doll[1,2], Sebastian Virreira Winter[1,2], Jakob M. Bader[1], Niklas Köhler[4], Fabian Theis[4,5], Alberto Santos[3,6] & Matthias Mann[1,3]✉

Proteins carry out the vast majority of functions in all biological domains, but for technological reasons their large-scale investigation has lagged behind the study of genomes. Since the first essentially complete eukaryotic proteome was reported[1], advances in mass-spectrometry-based proteomics[2] have enabled increasingly comprehensive identification and quantification of the human proteome[3–6]. However, there have been few comparisons across species[7,8], in stark contrast with genomics initiatives[9]. Here we use an advanced proteomics workflow—in which the peptide separation step is performed by a microstructured and extremely reproducible chromatographic system—for the in-depth study of 100 taxonomically diverse organisms. With two million peptide and 340,000 stringent protein identifications obtained in a standardized manner, we double the number of proteins with solid experimental evidence known to the scientific community. The data also provide a large-scale case study for sequence-based machine learning, as we demonstrate by experimentally confirming the predicted properties of peptides from *Bacteroides uniformis*. Our results offer a comparative view of the functional organization of organisms across the entire evolutionary range. A remarkably high fraction of the total proteome mass in all kingdoms is dedicated to protein homeostasis and folding, highlighting the biological challenge of maintaining protein structure in all branches of life. Likewise, a universally high fraction is involved in supplying energy resources, although these pathways range from photosynthesis through iron sulfur metabolism to carbohydrate metabolism. Generally, however, proteins and proteomes are remarkably diverse between organisms, and they can readily be explored and functionally compared at www.proteomesoflife.org.

To collect a diverse set of representative organisms across the tree of life, we considered the availability of assembled genome sequences and the accessibility of cultured or tissue material, and included common model organisms for comparison. This resulted in 19 archaea, 49 bacteria and 32 eukaryotes—a total of 100 different species (Fig. 1a, b). We also added 14 viruses (Supplementary Table 1).

To obtain the proteomes of these extremely different biomaterials, we tested a number of extraction protocols and found that the in-StageTip (iST) protocol[10] was most universally applicable and allowed automated and highly reproducible sample preparation. We incorporated the latest advances into our workflow for high-resolution bottom-up proteomics, and implemented a recently developed chip-based method[11] (Fig. 1c–e). C_{18}-covered beads are replaced by a uniformly ordered and statically fixed micrometre-sized pillar structure[12] (Fig. 1d), leading to 2.5-fold improvements in coefficients of variation for peptide retention times and high interlaboratory reproducibility (Extended Data Figs. 1, 2a). For all prokaryotes we performed single-run mass spectrometry (MS) analyses, whereas we used a loss-less prefractionator[13] for the more complex eukaryotic samples.

We reasoned that our chip-based chromatographic method, combined with the very large data set of more than two million unique peptides, should be well suited to deep learning algorithms, which have recently been shown to be applicable to MS-based proteomics[14–16] (Extended Data Fig. 3). We developed a long short-term memory (LSTM) deep learning model with an interpretable attention layer to precisely predict chromatographic retention times, achieving a Pearson correlation of 0.990 (Extended Data Figs. 2b, 4). To test the model on a completely unknown proteome, we instructed the mass spectrometer to sequence peptides from *B. uniformis*, *Bacillus megaterium* or *Enterobacter aerogenes* only if they eluted in a narrow band around the retention times predicted by deep learning. This resulted in only slightly diminished proteome depths (at least 88% on the protein level), showing that these peptide properties were successfully modelled in silico (Fig. 2).

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [2]OmicEra Diagnostics GmbH, Planegg, Germany. [3]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [4]Helmholtz Zentrum München–German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Munich, Germany. [5]Technical University of Munich, Department of Mathematics, Garching, Germany. [6]Li-Ka Shing Big Data Institute, University of Oxford, Oxford, UK. [7]These authors contributed equally: Johannes B. Müller, Philipp E. Geyer. ✉e-mail: mmann@biochem.mpg.de

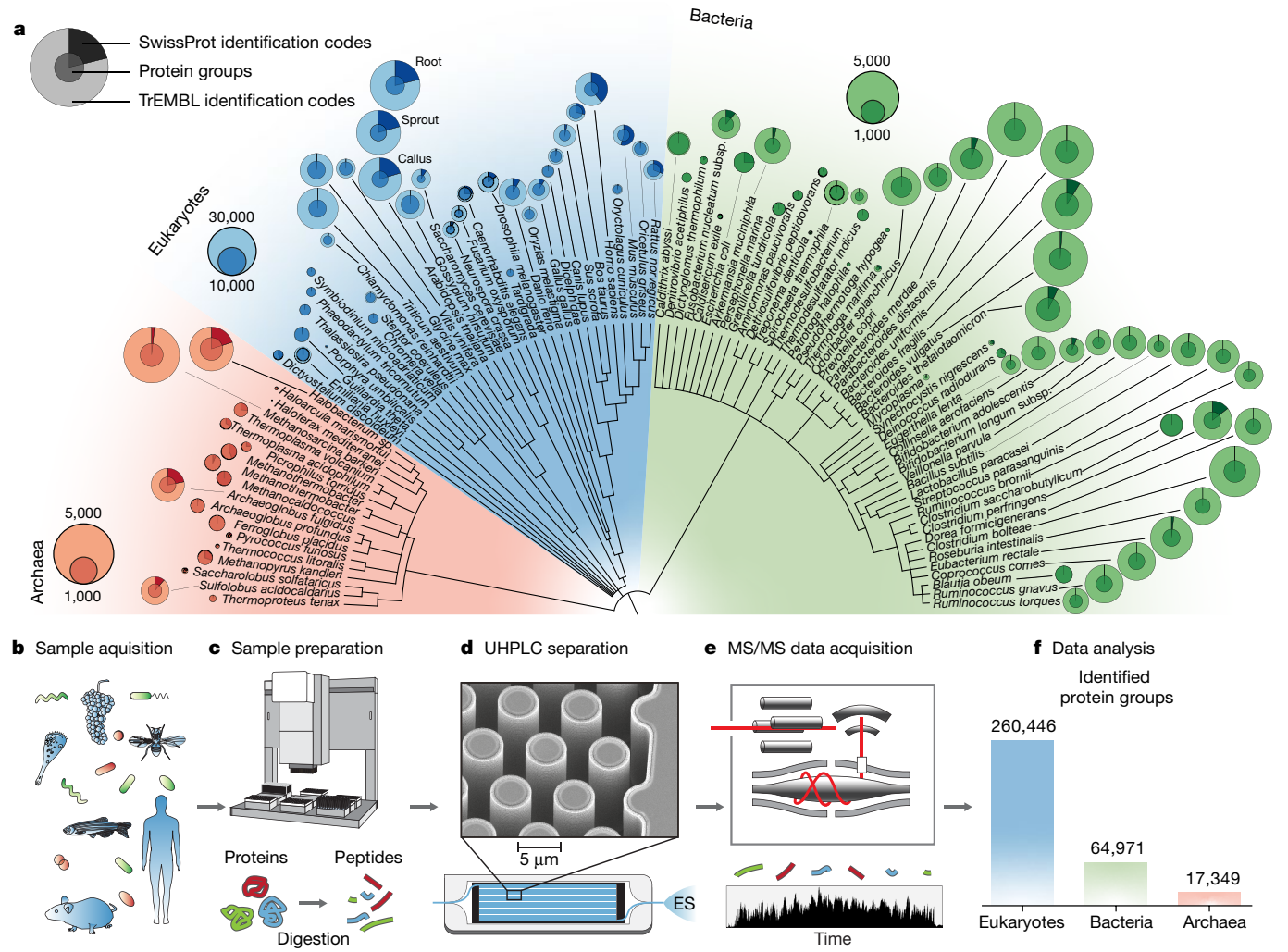


Fig. 1 | Collection of organism samples across the tree of life, and integration of the proteomic workflow. a, All organisms used herein were ordered and ranked on the basis of National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov) taxonomy. Pie charts refer to the numbers of protein groups (proteins distinguishable by their identified peptides) and to database protein entries found here. **b**, **c**, The acquired samples were subjected to protein extraction and digestion into peptides for sample preparation. **d**, Peptides were separated using a silica-chip-based micropillar array column (µPAC) with etched pillar structures that are coated with C_{18}. UHPLC, ultra-high performance liquid chromatography. The magnification shows a scanning electron microscopy image of the pillar structures (adapted with permission from PharmaFluidics). **e**, Peptides were ionized by electrospray (ES) and analysed in a high-resolution mass spectrometer. **f**, Numbers of identified proteins across the three superkingdoms.

Across the 100 organisms, we identified 349,164 proteins that were distinguishable by their identified peptides (Supplementary Table 2). These protein groups covered 1,136,558 entries, 93% of which were from TrEMBL—the section of the UniProt database (https://www.uniprot.org) that contains protein sequences predicted from genomes[17] (Fig. 1 and Extended Data Fig. 5). Because we have statistically significant evidence for the existence and correctness of our MS-derived peptide sequences, our data greatly increase the number of experimentally verified proteins, especially in bacteria and archaea. Contrary to our expectations, even well-studied model organisms still contributed many previously unknown proteins. The current Swiss-Prot database (version 2019_03, reviewed section of UniProt; see Methods) encompasses 559,634 experimentally verified proteins from all species. After taking into account proteins that have been described previously in the PRIDE/ProteomeXchange repository (https://www.ebi.ac.uk/pride/archive/), our additional 803,686 proteins more than double the number of proteins with experimental evidence.

To check the depth of proteome coverage, we inspected identifications for model organisms. With more than 5,000 identified protein groups in the yeast *Saccharomyces cerevisiae*, 9,000 in the zebrafish *Danio rerio* and 11,000 in the cotton plant *Gossypium hirsutum*, we obtained an even higher depth in comparison to previous large-scale efforts that focused on individual organisms. In prokaryotes we identified about half of all predicted genes at the protein level, representing a large fraction of the total proteome expressed in a single condition. However, this is less than the coverage obtained in several dedicated studies that used fractionation in these organisms and investigated different conditions. Eukaryotes generally have larger genomes and we identified correspondingly higher numbers of proteins (Fig. 1a). For instance, in a single human cell line, we identified 9,500 protein groups in our standardized workflow—a large proportion of the expressed proteome[6]—whereas 14 cell lines yielded 12,005 protein groups (Supplementary Table 4). Several species had very low proteome coverages. As the MS data were of similar quality in most of these cases (Supplementary Table 5), but the identification rates were low, we attribute

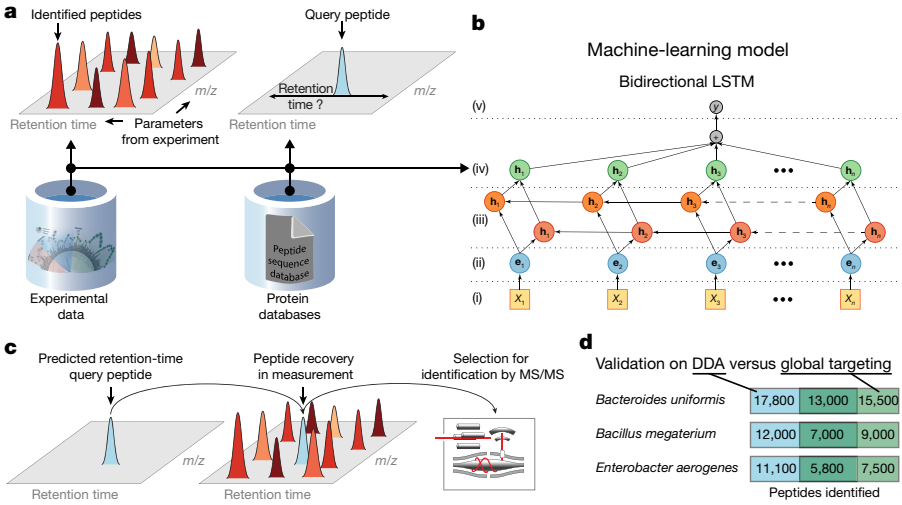


Fig. 2 | Application of a deep learning model to predict peptide retention times for liquid chromatography with tandem mass spectrometry (LC-MS/MS) measurements. a, The data used as inputs for retention time predictions are: left, our experimental data (from Fig. 1a), yielding retention time information on 2 million sequence-unique peptides from 100 organisms; and right, a list of query peptides with unknown retention times derived from a protein database. **b**, Bidirectional LSTM model with attention layer: (i), amino-acid sequence input (x_n); (ii), vectorization of amino-acid information for processing (yielding $\mathbf{e}_n$); (iii), generation of bidirectional LSTM layers ($\mathbf{h}_n$); (iv), attention-based reduction to fixed-length peptide-feature vector ($\mathbf{h}_n$); (v), prediction of retention time (y). **c**, Principle of the global targeting approach displayed for a single peptide: the instrument is set to select the peptide m/z peak for MS/MS identification if it is observed in a narrow retention time window predicted by deep learning. **d**, Application of the 'blind global targeting procedure' to all peptides of three previously unanalysed organisms resulted in the successful detection of predicted peptides in the organism samples. DDA, data-dependent acquisition.

the low proteome coverage to poor genome annotation or proteome prediction, which our data could help to improve through proteogenomics approaches.

In contrast to genomics and transcriptomics, proteomics data allow the direct estimation of the end product of gene expression[18]. We used label-free quantification in MaxQuant to estimate fractional protein intensities across multiple species[19]. Next, we asked how the proteins are distributed across the abundance range of the different organisms, and calculated the number of proteins that contribute to 90% of the total protein amount. The average was 1,546 proteins in eukaryotes, 306 in bacteria and 262 in archaea (Fig. 3a and Extended Data Figs. 6, 7). We used protein homology to enable the quantitative comparison of protein levels between the different organisms. Homology inference is a challenging bioinformatics problem, especially in poorly annotated organisms[20]. To perform the comparison across the studied species, we used high-quality homology prediction from Evolutionary Genealogy of Genes: Non-Supervised Orthologous Groups (EggNOG 5.0)[21]—a database of orthologous groups and functional annotations. We connected our quantitatively determined proteins and corresponding peptides with annotation and structural information data from various sources[17,22–24] in a graph database[25] yielding an explorable network structure with more than 8 million nodes (from proteins, peptides, gene ontology terms, and so on) and more than 53.8 million relationships between them (from homologies, associations, and so on) (Fig. 3b). The graph can be easily queried for any relationship between all of these nodes, as visualized for MS-identified homologues of two species (Fig. 3b). Here an abundant but uncharacterized protein from soybean (*Glycine max*) is linked to its counterpart in wine (*Vitis vinifera*), allowing direct comparison of MS identification, quantification and functional annotations. Similar queries can be performed for entire MS-characterized pathways, organelles or cell compartments. Co-varying pathways or gene ontology terms can also be explored, as well as their relationships to uncharacterized proteins (see www.proteomesoflife.org).

For instance, in soybean, the 11,208 quantified proteins covered more than five orders of magnitude (Fig. 3c) and had 1,763 annotated gene ontology terms. Applying a one-dimensional enrichment analysis to the annotated proteins[26] resulted in 734 statistically significantly enriched terms ($P < 0.05$) (Fig. 3d). Proteins linked to oxidation and reduction processes were the most abundant, reflecting the dominant roles of redox chemistry as a foundation for biochemical reactions such as glycolytic and carbohydrate metabolic processes (among the next most abundant categories). Apart from 'translation process', the most abundant gene ontology term of a biological process was 'protein folding', with an entire 3% of the protein mass. Altogether, functions dedicated to the life cycle of the proteome (translation, elongation, folding and proteolysis) made up a remarkable 10% of proteome mass in living organisms.

Conversely, certain classes of proteins were predominant only in specific branches of life (Extended Data Fig. 8). As expected, photosynthesis-related proteins were present only in photoautotrophic organisms such as plants, algae, protozoa or cyanobacteria (13 out of the 100 organisms) (Fig. 4 and Extended Data Fig. 9). Likewise, numerous functional associations can only be found within Bilateria or even Amniota. These mainly concern proteins associated with differentiation and tissue formation, higher intracellular spatial organization and well-described but subtaxonomy-specific signalling cascades. As expected, protein phosphorylation is predominantly but not exclusively present in eukaryotes. The bacteria and archaea both encompass organisms using this process (for instance in phosphorelay signalling), yet the proportion of the proteome mass involved in it is an order of magnitude lower in these organisms than in eukaryotes.

Much of proteome regulation is accomplished by post-translational modifications, which are typically investigated using specific enrichment protocols followed by MS analysis. However, even our nonenriched workflow in combination with the pFind tool[27] yielded a very large number of peptides with post-translational modifications for which the numbers of modified peptides were proportional to the size of the identified proteome (Extended Data Fig. 10). For instance, we found 29,426 serine phosphorylation sites, almost exclusively in eukaryotes, and 2,862 phosphotyrosine sites were largely restricted to ophistokonts (Supplementary Table 3).

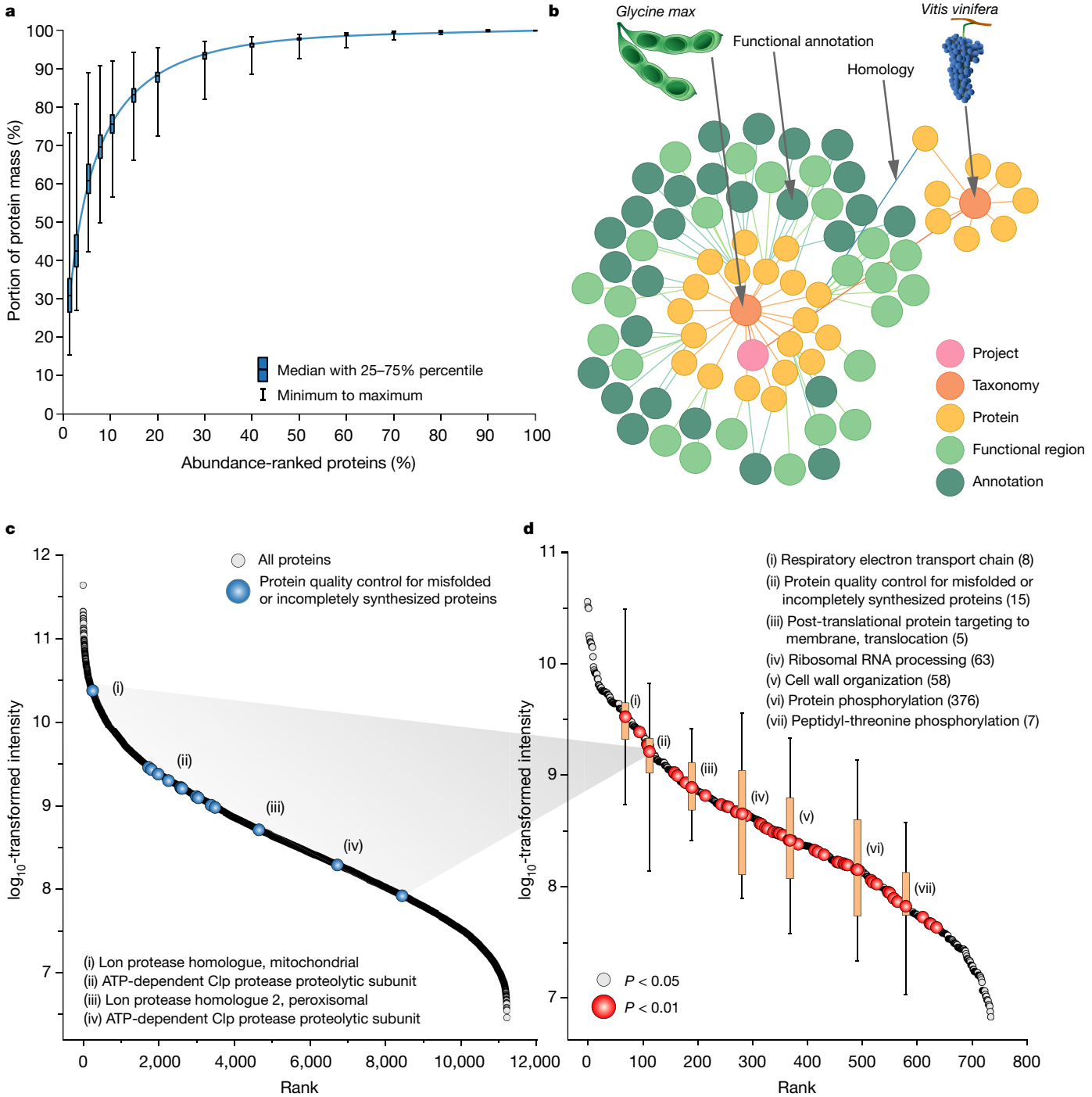


Fig. 3 | Organism-resolved integration of proteome data into a global analysis. a, Cumulative protein intensities (ranked by abundance; *x* axis) and their contribution to total protein mass (*y* axis) across all organisms ($n = 100$ organisms). **b**, Exemplified structure from the data model of the graph database, illustrating the connection between two homologous proteins of *G. max* and *V. vinifera*, and related annotations. **c**, All quantified proteins from *G. max* are displayed, plotting their intensities against their rank in the dynamic range. All proteins for which the functions are associated with 'protein quality control for misfolded or incomplete synthesized proteins' are highlighted. **d**, Significantly enriched functions (grey circles, $P < 0.05$; red circles, $P < 0.01$) within the proteome of *G. max* (with seven specific examples) and their distribution across the dynamic range (sample sizes in parentheses; one-sided Mann–Whitney *U*-test to the mean functional expression level). Error bars represent minimum to maximum values, and boxes show 10–90% percentiles.

Overall, 38.4% of the identified proteins did not have any functional annotation for the biological processes, and interestingly this was true even for 22.9% of the 100 most highly abundant proteins of each species at the biological-process level, and for 10% when considering protein functional domains (Extended Data Fig. 7 and Supplementary Table 6). Thus, our data point to a very large number of highly expressed proteins without any functional annotation or sequence homology to proteins with known gene ontology terms. Exploration of this part of the 'dark proteome' would be attractive: these proteins may indicate essential but unique features in the evolutionary development of these organisms that may be of biological or biotechnological interest.

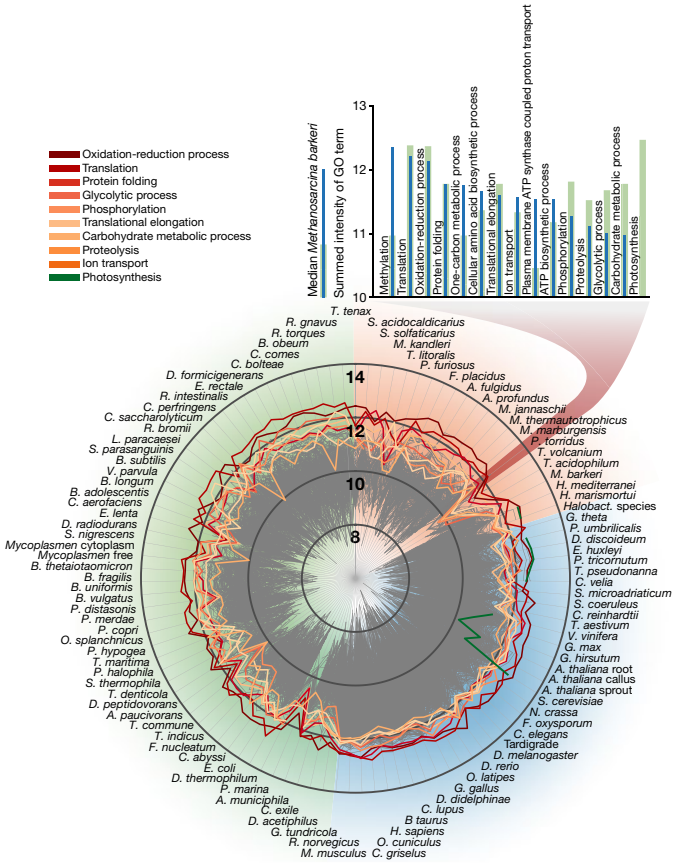


Fig. 4 | Global view of the expression levels of functional groups across the 100 organisms. The main diagram shows summed intensities for functional terms (grey lines), with the ten most abundant terms in all organisms colour-coded according to the key in the top left. The inset in the top right shows the most abundant gene ontology (GO) terms for the archaea *Methanosarcina barkeri* (blue lines), together with the median abundance of all 100 organisms for the displayed terms (green lines).

Advances in sequencing technology are now delivering the genome sequences of an exponentially increasing number of organisms, and we here made a first step towards a parallel scale-up of the characterization of proteomes. Sampling across the taxonomy of life, we created a large set of proteomes with high coverage of their expressed proteins. Label-free quantification values allow us to infer common and specialized biological functions and to compare them to close and distant relatives from all taxonomic levels. The data can be interactively explored at www.proteomesoflife.org.

Limitations of this study include the fact that we measured only selected cell types, tissues and biological states, and that the depth of proteome coverage is not yet comprehensive. Likewise, we have hardly touched upon the post-translational modification of proteins and their evolutionary diversity[28]. Ongoing improvements in MS-based proteomics—including more-refined abundance estimates[29], as well as entire streamlined workflows as described here—will substantially increase throughput in the future[2]. Given the cost effectiveness of proteomic measurements (marginal costs of less than $1,000 per species if its genome is available) and considering the wealth of novel data generated, we propose a community effort to explore many more organisms in different functional states. Integration with genomic, metabolomic and other data, together with incorporation of machine learning methods for species-specific libraries, would expand the systems-biological perspective beyond model organisms to the entire tree of life.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2402-x.

1. de Godoy, L. M. F. et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).
2. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
3. Nagaraj, N. et al. System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell. Proteomics* **11**, M111.013722 (2012).
4. Kim, M.-S. et al. A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
5. Wilhelm, M. et al. Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
6. Bekker-Jensen, D. B. et al. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Syst.* **4**, 587–599 (2017).
7. Weiss, M., Schrimpf, S., Hengartner, M. O., Lercher, M. J. & von Mering, C. Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics* **10**, 1297–1306 (2010).
8. Marx, H. et al. A proteomic atlas of the legume *Medicago truncatula* and its nitrogen-fixing endosymbiont *Sinorhizobium meliloti*. *Nat. Biotechnol.* **34**, 1198–1205 (2016).
9. Shendure, J. et al. DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017); correction *Nature* **568**, E11 (2019).
10. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
11. Geyer, P. E. et al. Plasma proteome profiling to assess human health and disease. *Cell Syst.* **2**, 185–195 (2016).
12. De Beeck, J. O. et al. Digging deeper into the human proteome: a novel nanoflow LCMS setup using micro pillar array columns (μPAC™). Preprint at *bioRxiv* https://doi.org/10.1101/472134 (2018).
13. Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics* **16**, 694–705 (2017).
14. Zhou, X.-X. et al. pDeep: predicting MS/MS spectra of peptides with deep learning. *Anal. Chem.* **89**, 12690–12697 (2017).
15. Tiwary, S. et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **16**, 519–525 (2019).
16. Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
17. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47** (D1), D506–D515 (2019).
18. Muñoz, J. & Heck, A. J. R. From the human genome to the human proteome. *Angew. Chem. Int. Edn* **53**, 10864–10866 (2014).
19. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
20. Altenhoff, A. M. et al. Standardized benchmarking in the quest for orthologs. *Nat. Methods* **13**, 425–430 (2016).
21. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47** (D1), D309–D314 (2019).
22. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47** (D1), D330–D338 (2019).
23. Geer, L. Y. et al. The NCBI BioSystems database. *Nucleic Acids Res.* **38**, D492–D496 (2010).
24. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47** (D1), D427–D432 (2019).
25. Santos, A. et al. Clinical knowledge graph integrates proteomics data into clinical decision-making. Preprint at bioRxiv https://doi.org/10.1101/2020.05.09.084897 (2020).
26. Cox, J. & Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **13** (Suppl 16), S12 (2012).
27. Chi, H. et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **36**, 1059–1061 (2018).
28. Zielinska, D. F., Gnad, F., Schropp, K., Wiśniewski, J. R. & Mann, M. Mapping N-glycosylation sites across seven evolutionarily distant species reveals a divergent substrate proteome despite a common core machinery. *Mol. Cell* **46**, 542–548 (2012).
29. Wiśniewski, J. R., Wegler, C. & Artursson, P. Multiple-enzyme-digestion strategy improves accuracy and sensitivity of label- and standard-free absolute quantification to a level that is achievable by analysis with stable isotope-labeled standard spiking. *J. Proteome Res.* **18**, 217–224 (2019).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Sample preparation

Organisms were obtained as stated in Supplementary Table 1. Cell lines were implicitly authenticated by MS and tested for mycoplasma contamination. The LLC-PK1 cell line was contaminated and mycoplasma contamination was harvested for analysis.

We carried out sample preparation according to the in-StageTip protocol[10] with an automated set-up on an Agilent Bravo liquid-handling platform as described[11]. In brief, samples were incubated in PreOmics lysis buffer (catalogue number P.O. 00001, PreOmics) for reduction of disulfide bridges, cysteine alkylation and protein denaturation at 95 °C for 10 min. Root and sprout parts of *Arabidopsis thaliana*, whole *Drosophila melanogaster* and leaves of *Porphyra umbilicalis* were ground in liquid nitrogen with a mortar and pestle beforehand. Samples were sonicated using a Bioruptor Plus from Diagenode (15 cycles, each of 30 s), and the protein concentration was measured using a tryptophan assay. In total, 200 µg of protein from each organism were further processed on the Agilent Bravo liquid-handling system by adding trypsin and LysC (at a 1:100 ratio of enzyme to sample protein, both in micrograms), mixing and incubating at 37 °C for 4 h.

We purified the peptides in consecutive steps according to the PreOmics iST protocol (www.preomics.com). After elution from the solid-phase extraction material, the peptides were completely dried using a SpeedVac centrifuge at 60 °C (Eppendorf, Concentrator Plus). Peptides were suspended in buffer A* (2% acetonitrile (v/v), 0.1% trifluoroacetic acid (v/v)) and sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510). Eukaryotes generally have larger numbers of genes than bacteria and archaea, resulting in a larger number of proteins and consequently of peptides. To reduce the complexity in the MS measurements, we separated eukaryotic peptide mixtures into eight fractions using the high-pH reversed-phase 'spider fractionator' as described[13].

UHPLC and mass spectrometry

We analysed the samples by applying LC-MS instrumentation, comprising an EASY-nLC 1200 ultrahigh-pressure system (Thermo Fisher Scientific) coupled to a Q Exactive HFX Orbitrap instrument[30] (Thermo Fisher Scientific) with a nano-electrospray ion source (Thermo Fisher Scientific).

For each analysis, 500 ng of purified peptides were separated on a 200 cm µPAC C_{18} microchip nano-LC column (PharmaFluidics). Peptides were loaded in buffer A*. To overcome the void volume of 10 µl, we applied a concentration gradient from 5% buffer B (0.1% formic acid (v/v), 80% acetonitrile (v/v)) to 10% buffer B coupled with a flow gradient from 750 nl min^{-1} to 300 nl min^{-1} for the first 15 min. Subsequently peptides were eluted with a linear gradient from 10% to 30% buffer B in 125 min at a constant flow rate of 300 nl min^{-1}. This was followed by a stepwise increase of buffer B to 60% in 5 min and to 95% buffer B in 5 min. Afterwards we applied a 5 min wash with 95% buffer B, followed by a 5 min decrease to 1% buffer B and a 20 min wash. We kept the column temperature constant at 50 °C by using an oven from Phoenix S&T (catalogue number PST-BPH-15). To avoid interference between the electrospray voltage and the µPAC chip column, we grounded the post-column connection, which was connected by a 20 cm long, 20 µm inner diameter fused silica post-column line to a New Objective Pico-Tip Emitter. This setup is further detailed in Extended Data Fig. 1b. The electrospray voltage was applied by connecting the mass spectrometer source output to the metal connection between the post-column sample line with an in-house-made clamp connection.

HPLC parameters were monitored in real time using SprayQC software[31]. MS data were acquired with a Top15 data-dependent MS/MS method. Target values for the full-scan MS spectra were 3×10^6 charges in the m/z range 300–1,650, with a maximum injection time of 20 ms and a resolution of 60,000 at m/z 200. Fragmentation of precursor ions was performed by higher-energy C-trap dissociation (HCD) with a normalized collision energy of 27 eV. MS/MS scans were performed at a resolution of 15,000 at m/z 200 with a target value of 1×10^5 and a maximum injection time of 28 ms. Dynamic exclusion was set to 30 s to avoid repeated sequencing of identical peptides.

Data analysis

MS raw files were analysed using MaxQuant software, version 1.6.1.13 (ref. [32]), and peptide lists were searched against their species-level UniProt FASTA databases. A contaminant database generated by the Andromeda search engine[33] was configured with cysteine carbamidomethylation as a fixed modification and amino-terminal acetylation and methionine oxidation as variable modifications. We set the false discovery rate (FDR) to 0.01 for protein and peptide levels, with a minimum length of seven amino acids for peptides. The FDR was determined by searching a reverse database. Enzyme specificity was set as carboxy-terminal to arginine and lysine as expected, using trypsin and LysC as proteases. A maximum of two missed cleavages was allowed. Peptide identification was performed in Andromeda with an initial precursor mass deviation of up to 7 ppm and a fragment mass deviation of 20 ppm. All proteins and peptides matching the reversed database were filtered out. All bioinformatics analyses were performed using Perseus[34] as well as standard analysis in Python version 3.6.4.

Machine learning model to predict retention times

To predict the retention times of peptides by machine learning, we isolated all detected peptide sequences, including modified peptides. For solvent-induced microshifts between runs, we corrected the detected retention times per peptide by the median shift of all peptides from one run to the median peptide retention time. This resulted in a total of 5,168,800 peptide sequences corresponding to 2,196,869 unique peptide sequences with a median retention time value for retention time prediction.

Our neural network architecture model takes a raw peptide sequence as input. Each amino acid was encoded into a 26-dimensional vector representation for processing using a one-hot encoding scheme, resulting in an $Lx26$ feature vector for a peptide with length L. This vector was connected to a two-layer bidirectional recurrent network with LSTM units with 500 hidden nodes each, which extract context-based features for each individual amino acid. This amino-acid-based feature embedding was reduced to a global 128-dimensional peptide-feature vector by an attention layer, which predicts the contribution of each individual amino-acid feature vector to the regression task. This peptide-feature vector was the input to a logistic regression layer, which regresses the expected retention time for the peptide sequence. The combination of recurrent layers with the attention layer allowed the model architecture to process peptide sequences with arbitrary lengths, but at the same time allow interpretability. The model was end-to-end trained on 2,125,113 peptides and validated on 54,490 holdout peptides. To validate the retention time prediction in vitro, we used the trained model to predict the peptide retention times of all tryptic peptides from *B. uniformis*, which were not included in the training set. We set the mass spectrometer to sequence only if the peptide eluted in a window of 1.4 s around the predicted retention time. This 'global targeting' was done using MaxQuant.life software (version 0.15)[35].

Graph database and cloud data-analysis notebook

To allow exploration of the MS experimental results, we developed a graph database (Neo4j: http://neo4j.com/, version 3.5.8, community edition) that collects all of the experimental data as well as homology and

functional annotations from different publicly available resources[17,21–24,36]. The implemented data model contains 11 different types of node and 14 types of link among the nodes; the data amount to 7,410,594 nodes and 35,517,979 relationships (5.02 GB). To populate the graph, flat files from source databases were downloaded and parsed to generate tab-delimited files comprising nodes and relationships, and standardized using selected terminologies and ontologies. The relationships collected in the database describe ontology structures (Directed Acyclic Graph relationships) and homology (orthology or paralogy) or functional associations (biological processes, functional regions, and so on). A version of the database is accessible at http://www.proteomesoflife.org.

The website gives access to interactive analyses implemented in Python (version 3.6), and uses Cypher as the query language (https://neo4j.com/developer/cypher-query-language/) (see also ref. [37]).

Data integration and comparison

We compared data in online proteomics repositories (PRIDE (https://www.ebi.ac.uk/pride/) and ProteomeXchange (http://www.proteomexchange.org)) with our data from 100 organisms, and downloaded either the provided protein tables or the raw files (Supplementary Table 6). We analysed the raw files with the same MaxQuant version and sequence files as used in our study. If identifiers other than UniProt identifiers were used, we applied the UniProt database to find the corresponding entries and to determine those proteins for which there was previous MS evidence.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The MS-based proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository and are available via ProteomeXchange with identifier PXD014877 and PXD019483.

Code availability

Custom computer code is available at https://github.com/MannLabs/proteomesoflife.

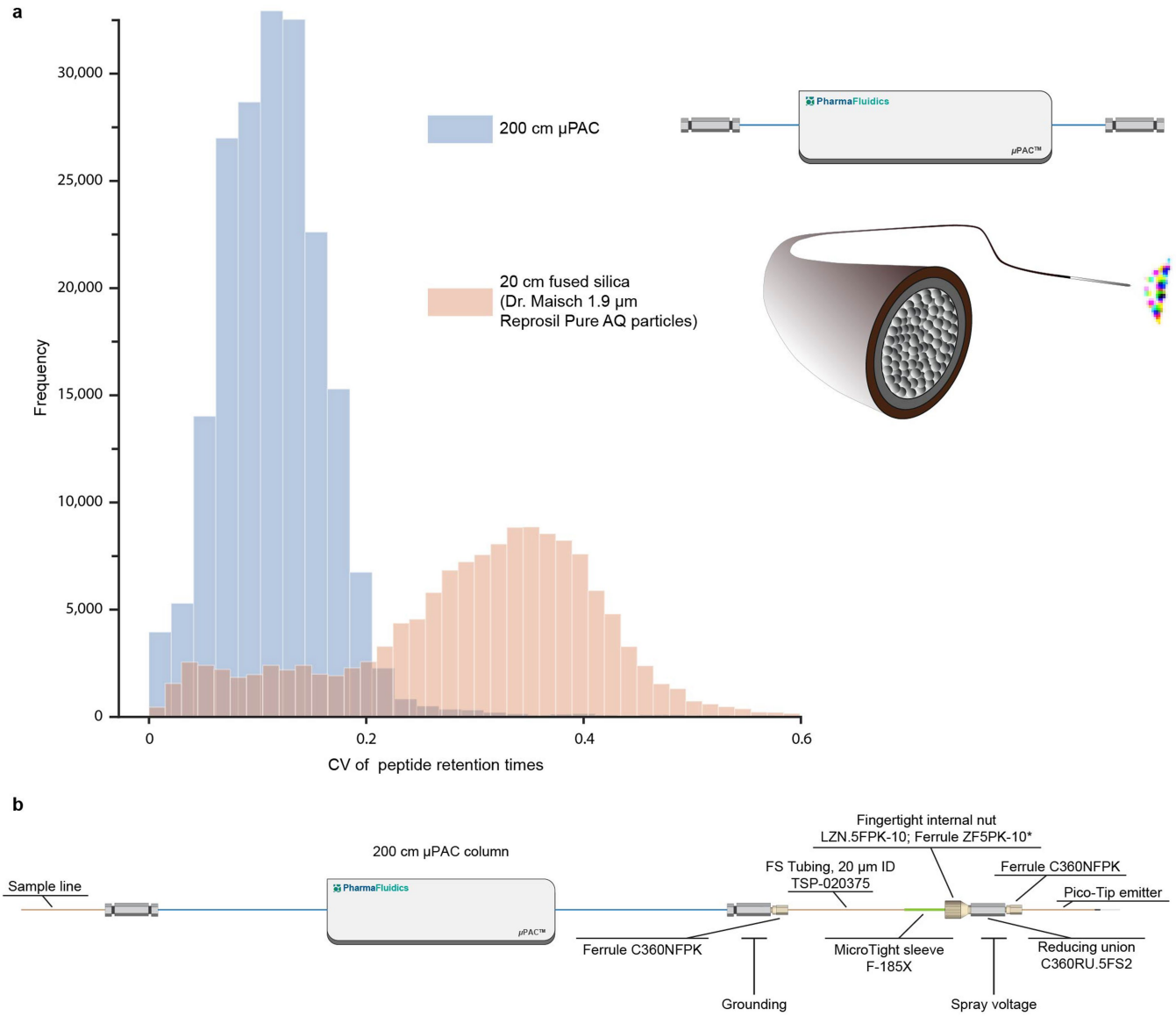
30. Kelstrup, C. D. et al. Performance evaluation of the Q Exactive HF-X for shotgun proteomics. *J. Proteome Res.* **17**, 727–738 (2018).
31. Scheltema, R. A. & Mann, M. SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J. Proteome Res.* **11**, 3458–3466 (2012).
32. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
33. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
34. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
35. Wichmann, C. et al. MaxQuant.Live enables global targeting of more than 25,000 peptides. *Mol. Cell. Proteomics* **18**, 982–994 (2019).
36. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47** (D1), D442–D450 (2019).
37. Perkel, J. M. Why Jupyter is data scientists' computational notebook of choice. *Nature* **563**, 145–146 (2018).

Acknowledgements We thank all members of the Proteomics and Signal Transduction Group and the Clinical Proteomics Group at the Max Planck Institute of Biochemistry, Martinsried, for help and discussions, and in particular I. Paron, C. Deiml, A. Strasser and B. Splettstoesser for technical assistance. We further thank the P. Bork group for supplying bacteria, the A. Pichlmair group for virus samples, F. Hosp for *A. thaliana*, I. Sinning for *Neurospora crassa* and the K.-P. Janssen group for cell line samples. Our work was partially supported by the Max Planck Society for the Advancement of Science, by the European Union's Horizon 2020 research and innovation program with the Microb-Predict project (grant 825694), by grants from the Novo Nordisk Foundation (NNF15CC0001 and NNF15OC0016692), and by the Deutsche Forschungsgemeinschaft (DFG) project 'Chemical proteomics inside us' (grant 412136960).

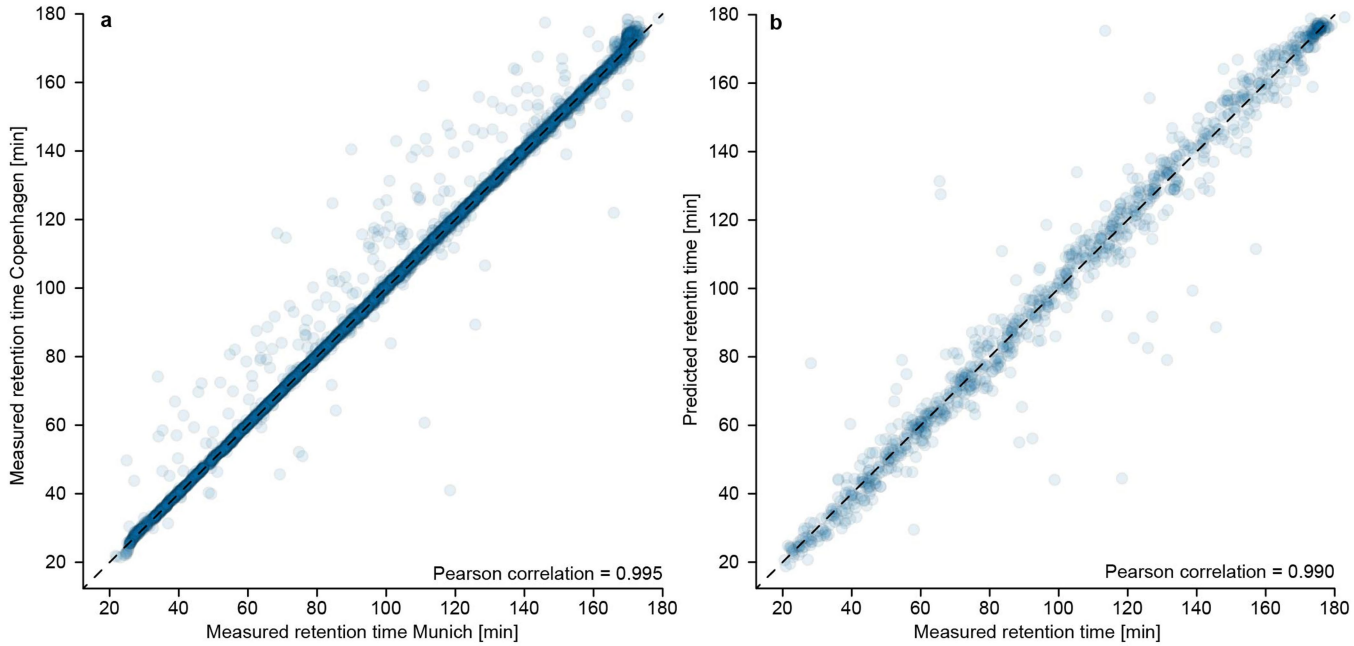
Author contributions J.B.M. and P.E.G. designed the experiments, performed and interpreted the MS-based proteomic analyses, carried out bioinformatics analyses and generated text and figures for the manuscript. P.V.T., S.D., S.V.W. and J.M.B. designed experiments and performed MS-based proteomics analyses. A.R.C. and A.S. integrated annotation data with proteomics data and implemented the Python code as well as graph-based structures. A.S. and M.O. implemented the web-accessible analyses. N.K., F.T. and M.T.S. carried out the machine learning analysis. M.M. supervised and guided the project, designed the experiments, interpreted MS-based proteomics data and wrote the manuscript.

Competing interests The authors declare no competing interests.

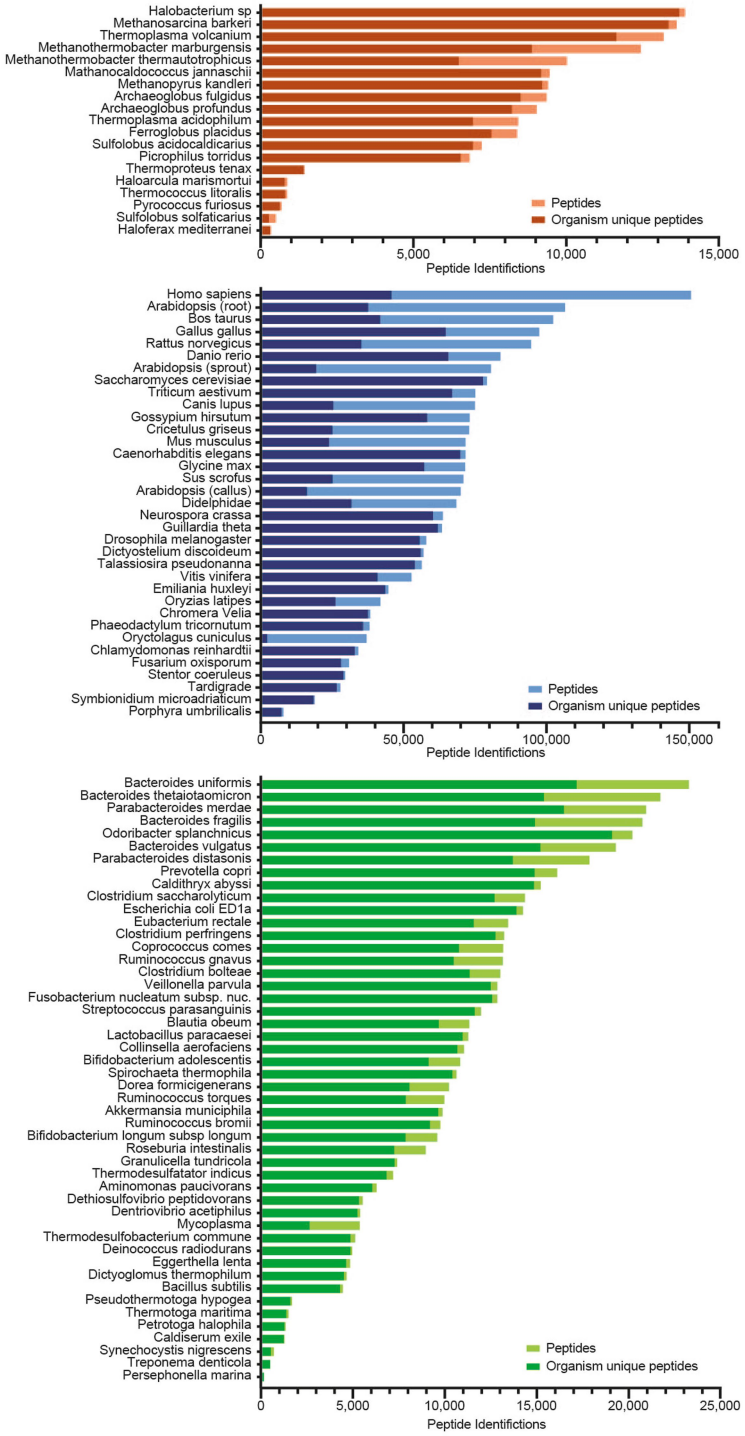
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2402-x.
Correspondence and requests for materials should be addressed to M.M.
Peer review information *Nature* thanks Joshua Coon, Vera van Noort and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Comparison of the peptide retention times obtained by a µPAC and a fused silica capillary column. a, The histograms illustrate the distribution of coefficients of variation (CVs) calculated from peptide retention times obtained by a µPAC and a fused silica capillary column. The CVs were calculated for peptides from 12 measurements of a HeLa cell digest on each column. **b**, All components, including lines, connectors, the column and the emitter, are displayed together with grounding and spray voltage connections. The pico tip emitter is from New Objective (catalogue number FS360-20-10-N-5-105CT).

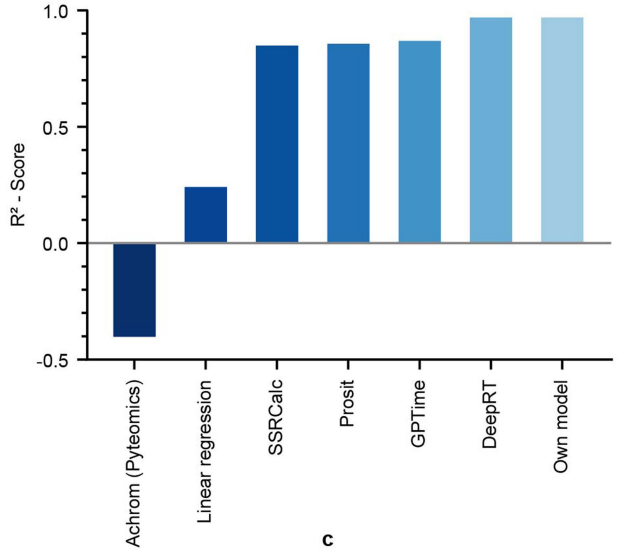


Extended Data Fig. 2 | Interlaboratory reproducibility and prediction of peptide retention time on the µPAC column. a, The ability to produce chip-based columns in a reproducible manner, coupled with the statically fixed micrometre-sized pillars, results in highly reproducible performance and interlaboratory transferability of the µPAC-based approach. Shown are the corrected retention times of an excerpt of 5,000 peptides from the 43,000 overlapping peptides measured in two different HeLa cell digests by our Munich and Copenhagen laboratories, resulting in a Pearson correlation coefficient of peptide retention times of 0.995. **b**, To validate our model for predicting peptide retention times, we plot an excerpt of 1,000 peptides from the complete test-set of 54,490 peptides, with experimentally determined values on the *x* axis and predicted values on the *y* axis. The Pearson's R^2 correlation value for the complete predicted peptide set is 0.99.



Extended Data Fig. 3 | Total numbers of identified peptides from 100 organisms across the tree of life. The peptides uniquely identified for a certain organism are colour-coded from peptides identified in multiple species. Orange, archaea; blue, eukaryotes; green, bacteria.

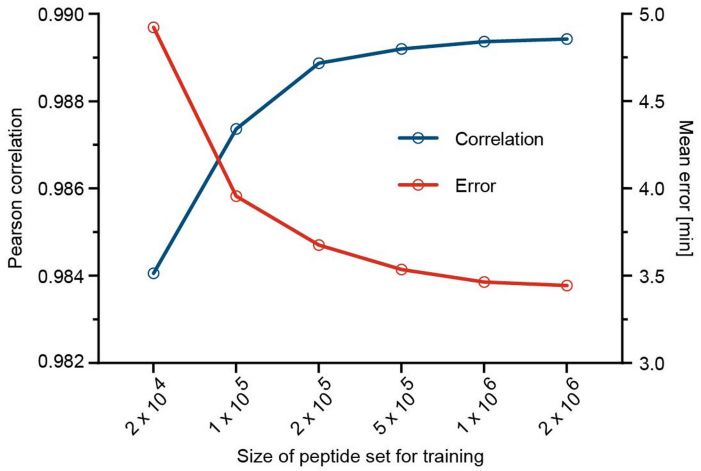
a



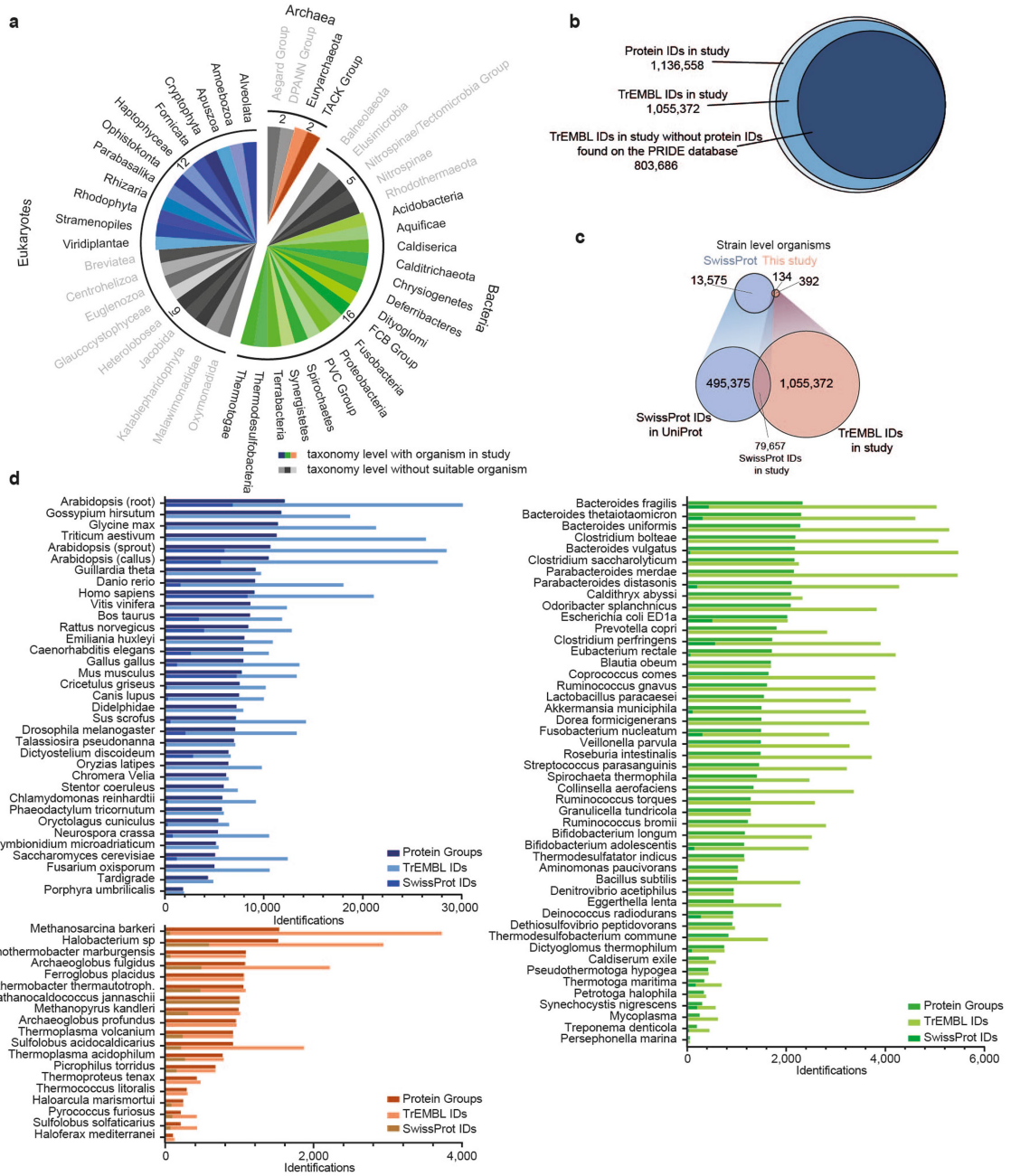
b

Model	R2-Score	Pearson	Training set	Test set
Achrom (Pyteomics)	-0.413	0.232	10000	10000
Linear Regression	0.253	0.503	1537808	659060
SSRCalc	0.860	0.928	1000	1000
Prosit	0.868	0.969	-	1000
GPTime	0.881	0.939	10000	10000
DeepRT	0.980	0.990	1537808	659060
Own model	0.980	0.990	2125113	50000

c

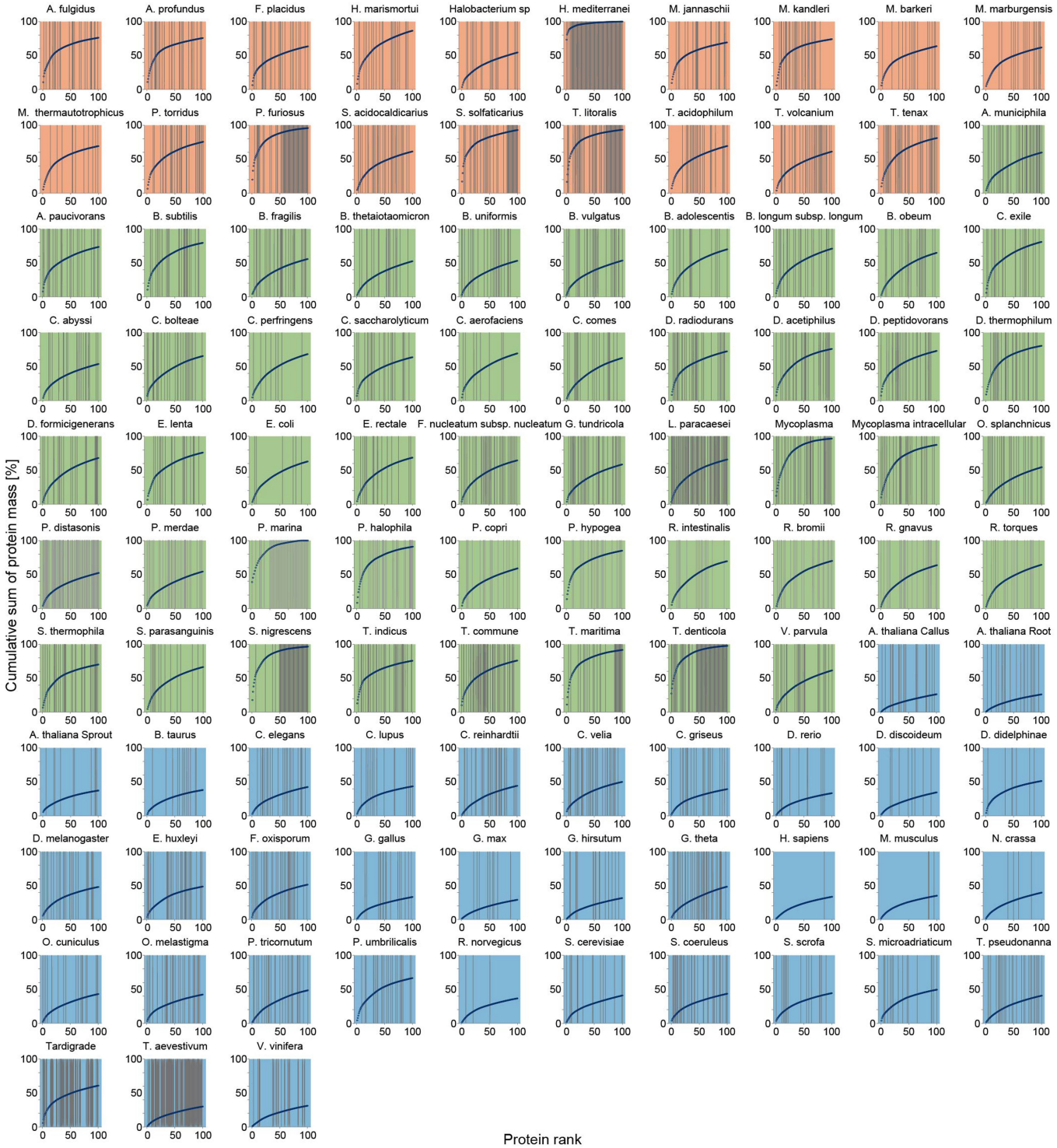


Extended Data Fig. 4 | Comparison and characterization of the LSTM model for predicting peptide retention times. a, Box plots comparing R^2 scores obtained from different models of peptide retention time, calculated from the linear regressions of correlations between the predicted test set to the measured peptide retention times. Sample sizes are shown in **b**. **b**, Table comparing the different models of peptide retention time. The training set was reduced in size (number of peptides included) in order to account for the exponentially growing calculation time of certain models. Statistics represent the linear regression of correlation from the predicted test set retention times to the measured retention times. **c**, Characterization of the LSTM model applied here for different sizes of training peptide set.

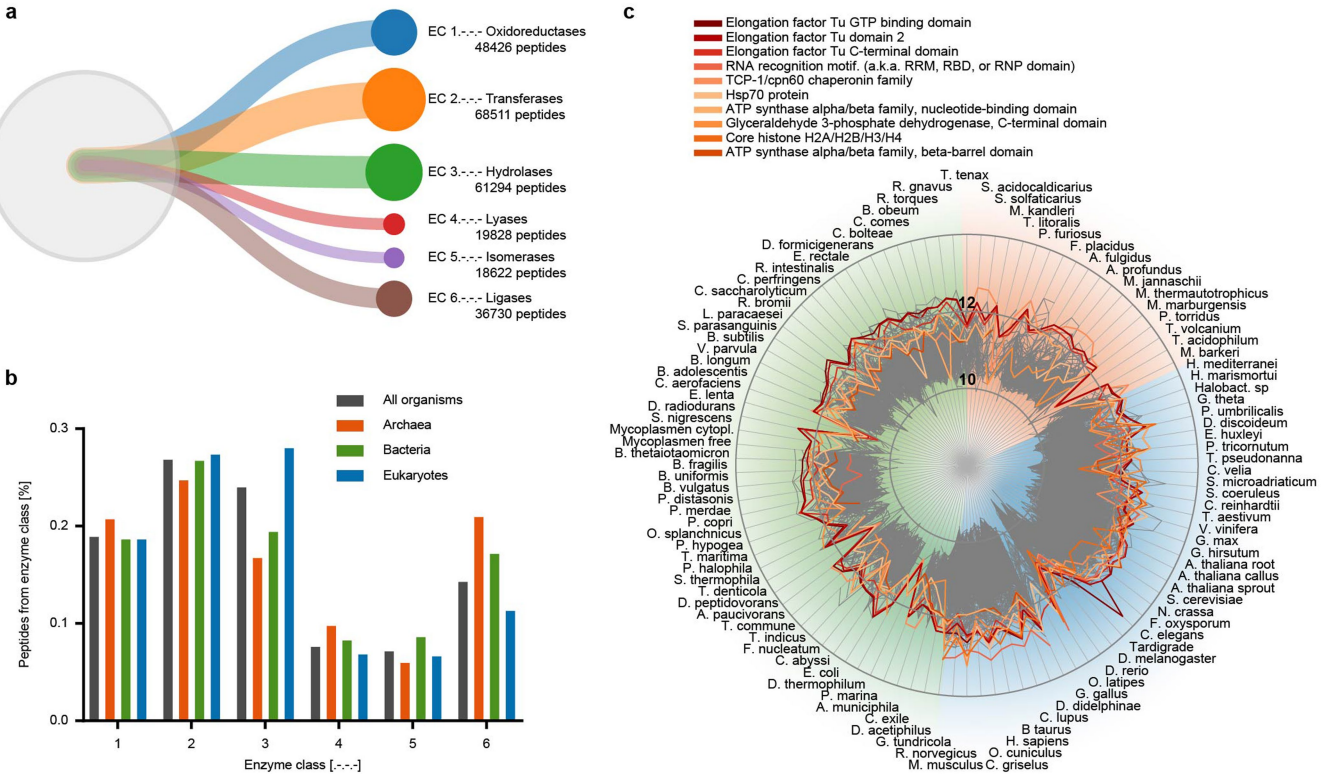


Extended Data Fig. 5 | Overview of our data set of 100 organisms across the tree of life. a, Illustration of all direct taxonomic levels below the superkingdom level that are covered by our data set. DPANN, Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota; FCB, Fibrobacteres, Chlorobi and Bacteroidetes; PCV, Planctomycetes, Chlamydiae and Verrucomicrobia; TACK, Thaumarchaeota, Crenarchaeota and Korarchaeota. **b**, Number of protein identification codes (IDs) in this study and their relation to TrEMBL IDs found in the PRIDE archive. **c**, Comparison of the Swiss-Prot database to the data set in this study with regards to organism and protein numbers. **d**, Numbers of identified protein groups and UniProt protein entries for all 100 organisms in our data set. The UniProt protein-entry identifications are colour-coded into Swiss-Prot (reviewed) and TrEMBL (predicted) entries.

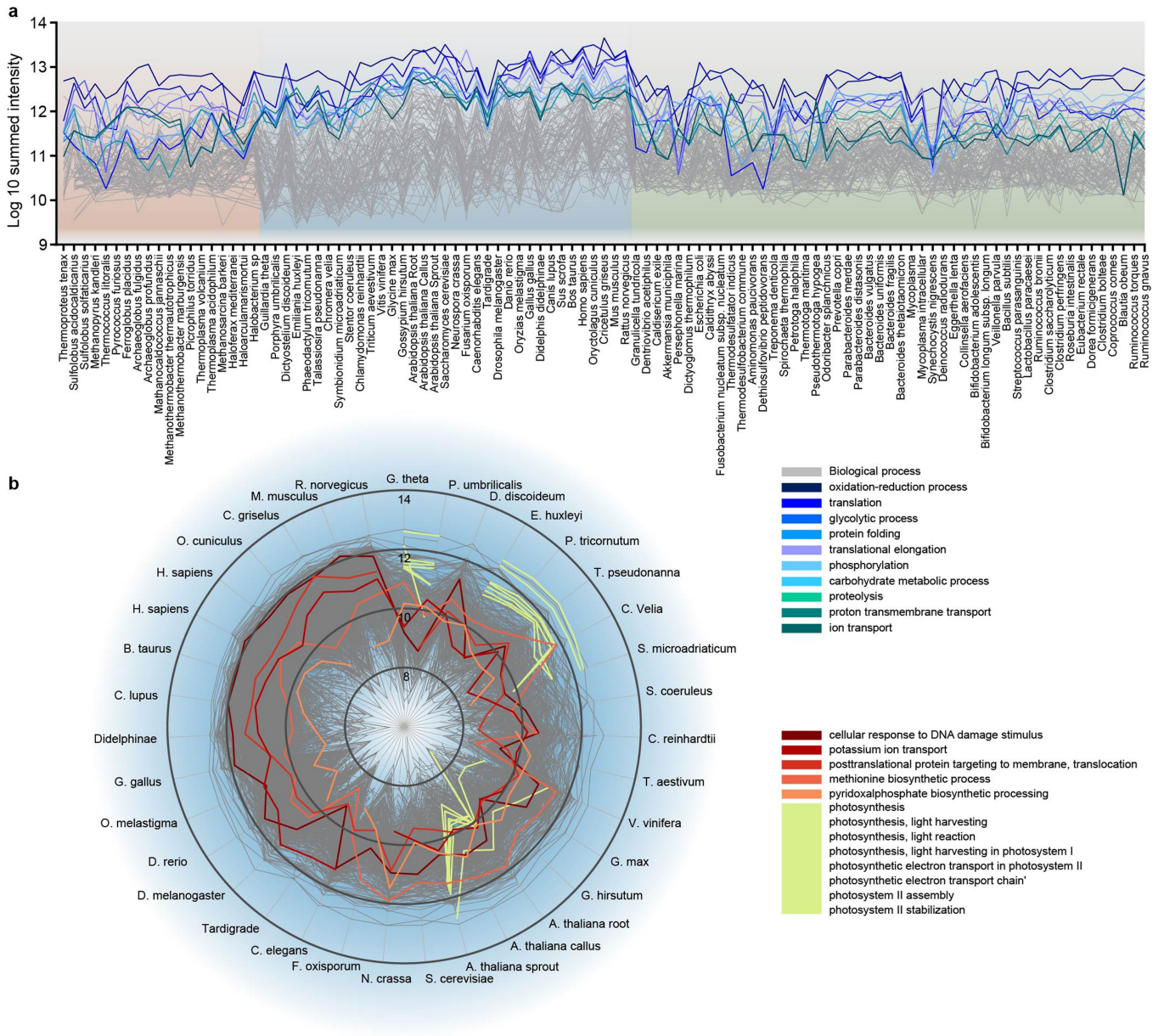
Extended Data Fig. 6 | Dynamic range curves for all organisms analysed here. Protein intensities are $\log_{10}$-scaled and plotted against the abundance rank of each protein.



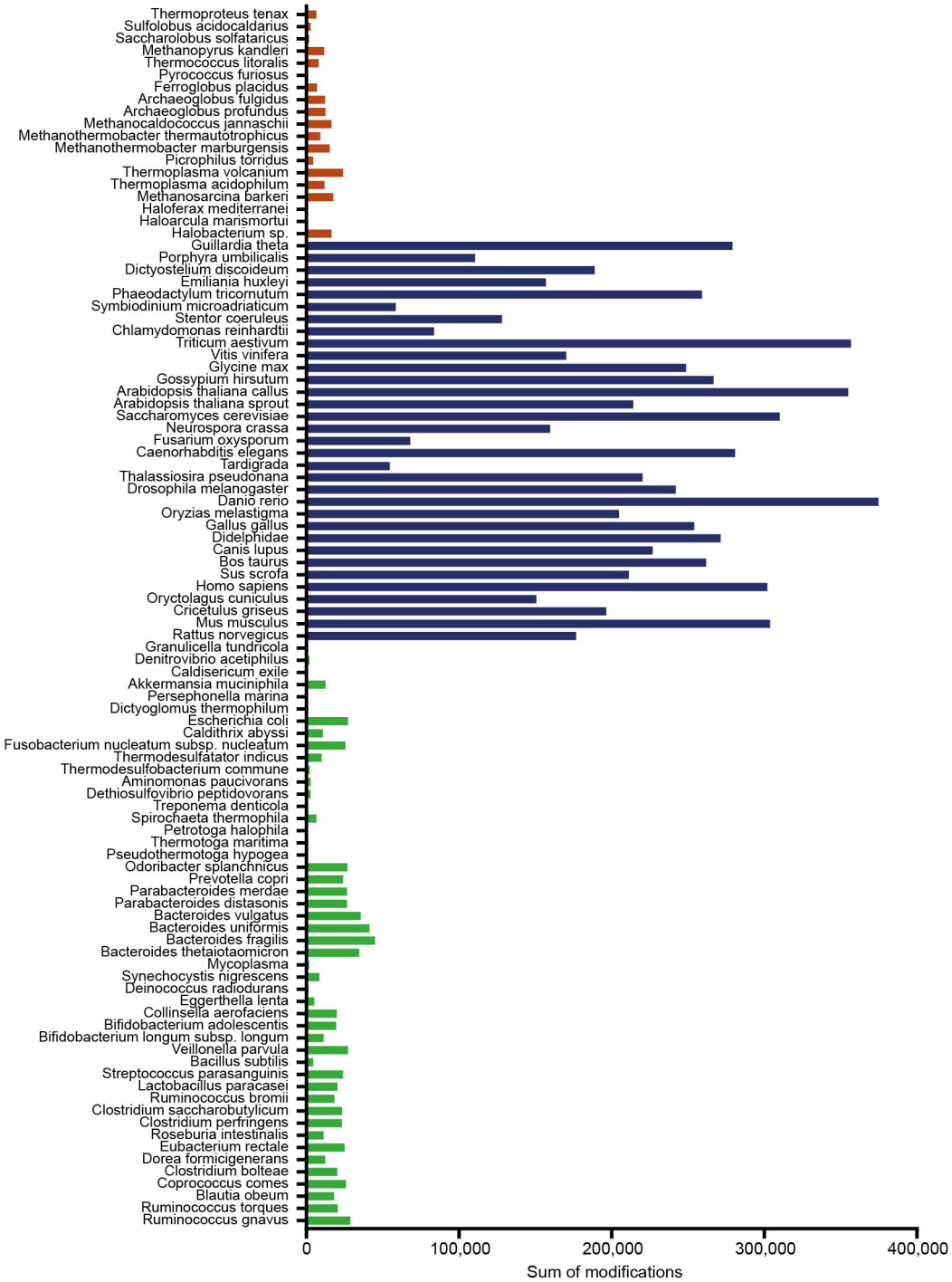
Extended Data Fig. 7 | Cumulative protein intensities for all organisms analysed here. On the *x* axis, proteins are ranked according to their abundance; the *y* axis shows the cumulative protein intensity. Proteins missing biological-process annotation are highlighted by grey lines in the background.



Extended Data Fig. 8 | Quantitative analysis of different enzyme classes and functional protein domains across the tree of life. a, We classified the contribution of peptides to the top 90% of protein mass within all 100 organisms according to the enzyme commission (EC) number, using the Unipept web-tool (https://unipept.ugent.be/). The alluvial plot illustrates the proportions of each enzyme class across all organisms in our study. **b**, Comparison of the three domains of life with respect to their normalized contribution of peptides to each enzyme class. **c**, Proteins that contribute to the top 90% of the protein mass within all 100 organisms studied herein were annotated according to their known functional protein domains, and the intensities for different functional domains of an organism were summed to display the most abundant functional protein domains across the tree of life. The intensity is displayed on a $\log_{10}$ scale.



Extended Data Fig. 9 | Quantitative analysis of specific biological processes across the tree of life. a, Linear display showing a global view of the expression levels of functional groups across the 100 organisms from Fig. 4. Summed intensities for functional terms are shown as grey lines, with the top ten most abundant terms in all organisms colour-coded according to the top key. **b**, Quantitative analysis of specific biological processes from the superkingdom of eukaryotes. Proteins were annotated with biological processes, and the intensities for each annotation term within an organisms were summed. Those biological processes that display differential expression across the superkingdom as well as photosynthetic processes are highlighted according to the bottom key.



Extended Data Fig. 10 | Modified peptides. Sum of modified peptides per organism, identified with pFind (http://pfind.ict.ac.cn/software/pFind3/index.html) and colour-coded for archaea (red), eukaryotes (blue) and bacteria (green).

nature research

Corresponding author(s): Prof Matthias Mann

Last updated by author(s): Feb 20, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The LC-MS data collection was performed with a Thermo Fisher Easy LC 1200, employing a Pharmafluidics 200 cm µPAC column using a Thermo Fisher Q Exactive HF-X mass spectrometer. Electrospray conditions were monitored by the SprayQC software.
Data analysis	LC-MS data processing was performed using MaxQuant software (version 1.6.1.13), employing the Andromeda search engine. Data analysis was done using Perseus software and custom code written in Python (version 3.6). For the interactive analyses a Neo4j (version 3.5.8 Community edition) was used to develope a graph database which is acces via Python (version3.6) and uses Cypher as query language.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

LC-MS raw files, FASTA files and MaxQuant output tables are available via the PRIDE partner repository via ProteomeXchange with identifier PXD014877.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	100 Organisms across the tree of life, thereof 19 archaea, 49 bacteria and 32 eukaryotes were analyzed in singlets by MS based proteomics.
Research sample	To collect a diverse set of representative organisms across the tree of life, we considered the availability of assembled genome sequences, the accessibility of cultured or tissue material and included common model organisms for comparison.
Sampling strategy	Cohort size of 100 organisms was chosen by a trade of between biological impact and data acquisition time. As a pioneer study in cross organism proteome comparison this represents a sufficient overview of the proteome of living organisms. Samples were prepared for bottom up proteomics with LC-MS, by denaturationg and lysing (boiling, sonication, grinding), reduction/ alkylation and tryptic digestion of proteins, followed by subsequent purification of peptides by desalting.
Data collection	Data was obtained by bottom up proteomic measurement of tryptic digested proteins with mass spectrometry. The experiments were carried out by the authors.
Timing and spatial scale	Data were collected from September 2018 to January 2019.
Data exclusions	No data were excluded.
Reproducibility	No repetitive measurements were conducted.
Randomization	Sample handling and measuring was done in batch processing. Fractions of the same samples were measured consecutively.
Blinding	Samples were not blinded due to batch processing.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	ATCC-CCL-22; ATCC-CCL-34; ATCC-CCL-61; ATCC-CRL-2050; ATCC-CRL-1840; 6C2; ATCC-CCL2; ATCC-CRL-3209; ATCC-CRL-6503; ATCC-CRL-1721; LLC-PK1
Authentication	Cell lines were not authenticated.
Mycoplasma contamination	One cell line was positive for Mycoplasma contamination which was used to generate a proteomics dataset for Mycoplasma.

Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell line was used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Drosophila melanogaster, strain CantonS with a mean age of 24 h and mixed sex were used. Caenorhabditis elegans N2 wild.type strain were used.
Wild animals	Study did not involve wild animals.
Field-collected samples	Study did not involve field-collected samples.
Ethics oversight	No ethical approvement was required.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Author Correction: Quantifying the dynamics of failure across science, startups and security

https://doi.org/10.1038/s41586-020-2325-6

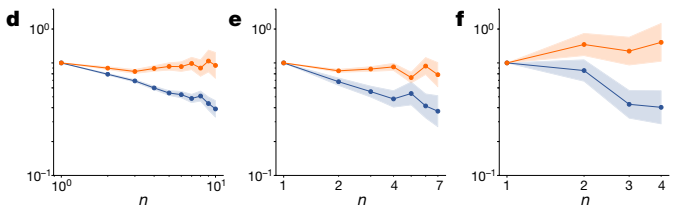
Correction to: *Nature* https://doi.org/10.1038/s41586-019-1725-y

Published online 30 October 2019

Yian Yin, Yang Wang, James A. Evans & Dashun Wang

In this Article, the '10^0' markers on the *y* axes of Fig. 3d–f were inadvertently misplaced, and the *y* axis labels were omitted. Figure 1 of this Amendment shows the original, incorrect panels and the corrected panels of Fig. 3d–f, for transparency to readers. The original Article has been corrected online.

Original Fig. 3d–f



Corrected Fig. 3d–f

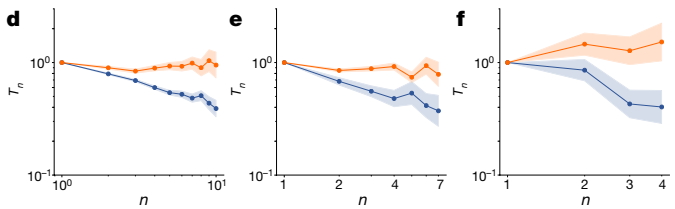


Fig. 1 | This figure displays the corrected and the incorrect published Fig. 3d–f of the original Article.

Author Correction: A conserved dendritic-cell regulatory program limits antitumour immunity

https://doi.org/10.1038/s41586-020-2326-5

Correction to: *Nature* https://doi.org/10.1038/s41586-020-2134-y

Published online 25 March 2020


Check for updates

Barbara Maier, Andrew M. Leader, Steven T. Chen, Navpreet Tung, Christie Chang, Jessica LeBerichel, Aleksey Chudnovskiy, Shrisha Maskey, Laura Walker, John P. Finnigan, Margaret E. Kirkling, Boris Reizis, Sourav Ghosh, Natalie Roy D'Amore, Nina Bhardwaj, Carla V. Rothlin, Andrea Wolf, Raja Flores, Thomas Marron, Adeeb H. Rahman, Ephraim Kenigsberg, Brian D. Brown & Miriam Merad

In this Article, Supplementary Tables 1 and 2 were originally not uploaded online; these files have now been added to the original Article.

Publisher Correction: Patterns and trends of Northern Hemisphere snow mass from 1980 to 2018

https://doi.org/10.1038/s41586-020-2416-4

Correction to: *Nature* https://www.nature.com/articles/s41586-020-2258-0

Published online 20 May 2020


Check for updates

Jouni Pulliainen, Kari Luojus, Chris Derksen, Lawrence Mudryk, Juha Lemmetyinen, Miia Salminen, Jaakko Ikonen, Matias Takala, Juval Cohen, Tuomo Smolander & Johannes Norberg

In this Article, owing to an error during the production process, the corresponding author was incorrectly indicated to be Johannes Norberg rather than Jouni Pulliainen. The original Article has been corrected online.

Work

Your story

Send your careers story to: naturecareerseditor@nature.com

GETTY

HOW TO BEAT ISOLATION? FEEDBACK GROUPS FIT THE BILL

Researchers share tips for supporting research and life from a distance.

By A. R. Siders, Cassandra M. Brooks, Amanda E. Cravens, Rebecca L. Nelson, Dan R. Reineman & Nicola Ulibarri

We are an interdisciplinary group of scientists who have met online each week over the past eight years to provide one another with feedback and support in our careers. We have built a virtual community that has enriched our lives as people and as scholars. As the coronavirus pandemic threatens to isolate researchers, our experience might be of value. The pandemic is putting many people through a difficult time, but it might be an opportunity to create networks that will provide immediate emotional support and long-term research benefits. We hope our advice on forming and maintaining a virtual feedback group will inspire others searching for ways to fight isolation and build communities.

Our project started as a dissertation-writing group while we were doing graduate work at Stanford University in California. It endured because, as we began our careers as academics and government scientists, we discovered that we had as much – if not more – need for one another's reactions and support. We found that feedback, especially that given early in the research process, was crucial to our scholarly success and was largely missing from available networks.

In our virtual meetings, one of us shares something – a paper, outline, study design, syllabus, grant proposal, preliminary data or mere seed of an idea – and receives comments from the group. Our emphasis of feedback on content and its scientific underpinnings distinguishes our group from those focused on writing, which prioritize output or editing. Nor are we a standard research group: although we routinely provide contributions to one another's academic works and acknowledge members in our papers, our purpose is to support individual research agendas and career progress by providing rapid, frequent and multiple-stage commentary from a group of respected peers in a safe environment.

This type of feedback has many benefits. It does not replace the need for field- or organization-specific advice from experts in our disciplines or from mentors at our institutions. But, by enabling us to engage early in the research process and with people outside our discipline, our group offers benefits that those sources might not. Receiving advice early in the research and teaching process improves quality in ways that late-stage feedback from peer review or conference presentations does not. Engagement with interdisciplinary colleagues introduces us to new literature and methods, enabling us to be better collaborators and mentors.

Giving and receiving feedback improves our contributions to journal peer review and our ability to advise students. All members learn from everyone else's experiences, accelerating exposure to scholarly and professional activities and giving us confidence and insight. One member has provided feedback on more than 40 articles and 2 book proposals in the group — which exceeds the norm for her position as a first-year faculty member. This has given her the confidence to propose her own book and take on a journal-editor position.

Belonging to a supportive group in which we can present challenging ideas and receive advice without judgement enables us to step out of our comfort zones. This helps us, for example, to advance our fields by employing new methods, and to maintain work–life boundaries by saying no to enticing but unnecessary opportunities. And our support has helped members to pursue creative projects: one of us has published work using geographic maps to explore legal concepts.

From our experience, we have some tips for beginning such a group:

Choose respected peers from outside your 'normal' groups

Each member must respect the contributions of every member. It can be helpful for some participants to know one another personally — by sharing a common graduate program, as we did, or by belonging to the same professional or scholarly community, for example — but this is not essential. The group will meet online, making it an excellent space to expand your network by including colleagues from outside your university or usual collaboration circles. Including someone from a conference whose work is tangential to your own or the author of a paper in a field slightly outside yours can be good starting points.

Find disciplinary breadth and common ground

Having members from different disciplines has exposed all of us to new ideas and methods, although there are limits to how much interdisciplinarity is helpful. Our members come from environmental sciences, geography, law, urban planning, environmental planning and public policy, so we vary widely in our methods, research interests and where we work; however, we share an interest in environmental governance and water. This gives us common ground and a shared language that helps to ensure our comments are informed and relevant.

Meet frequently and provide feedback at early stages

Unlike a writing group that aims to polish a nearly final product, feedback groups can mould nascent concepts. This helps members to filter promising ideas from unpromising ones, provides iterative commentary throughout the research process and encourages submission of rough concepts to enable prototyping of ideas.

Create a safe space

Feedback groups are most useful when members can be vulnerable and present a challenging idea or can offer constructive criticism without fear of reprisal. We engage from a place of respect: we direct our comments at the work, not the person. We balance negative and positive comments, and all feedback is presented constructively. Furthermore, we belong to different institutions and disciplines, so there is no 'career risk'. We are not on each other's tenure or promotion boards; we do not review one another's journal publications. Thus, we can ask questions that we might not raise in our own departments or organizations.

"For some, it might evolve into something lasting, with benefits that extend far beyond the current crisis."

Be flexible, but take the commitment seriously

For some, this group could just be a stop-gap, perhaps to address social isolation due to the coronavirus pandemic. Even if this is the case, we encourage you to embrace the potential of continuing the group afterwards. Participating over years builds trust and creates reciprocity. This enables members to share ideas that they are less confident about, and to ask more of each other, such as a quick response on a time-sensitive issue or a detailed review of a long piece of work. Each person knows the others have asked or will ask the same in return. Trust and reciprocity do not instantly materialize; they must be cultivated.

Our group has evolved over time, as have our 'rules'. Since we formed, members have moved between countries, states and jobs; we have got married, had kids, overcome illnesses and travelled. Flexibility helps us to address the realities of our changing lives while remaining committed to the group. We began with strict rules about weekly in-person attendance, and we used guidelines such as these to discuss expectations about how frequently we would meet, what the attendance policy would be and what types of feedback we would provide. Gradually, we relaxed these rules and transitioned to a weekly video conference with regular, if not compulsory, attendance. Members occasionally sit out for months at a time for career or family reasons. We developed trust and commitment to each other that rendered strict rules unnecessary.

Our technology and scheduling have also changed. Our members live in four time zones spanning a ten-hour time difference (from the east coast of the United States to the east coast of Australia). At the beginning of each semester, we review our teaching schedules and commitments, and use online scheduling platforms to find a meeting time that works for everyone for that semester. We have learnt that, no matter how well we plan, something will come up, such as daylight savings occurring on different days in different countries. Rather than let such hiccups derail our group, we have learnt to be flexible. We provide feedback through e-mail, change our presentation rota or simply skip a week. We are conscious that the group's goal is to provide support, not to create another must-do task.

For many scientists, a short-term group could help them to sustain their research practice and endure emotional strains during this uncertain time. But for some, it might evolve into something lasting, with benefits that extend far beyond the current crisis. The group has improved the quality of our scholarship and helped us to navigate challenges in our professional lives. Perhaps even more importantly, it has helped us to build a community and maintain work–life balance, which makes us happier academics and more well-rounded scholars and humans.

A. R. Siders is an assistant professor of public policy and geography at the University of Delaware, Newark. **Cassandra M. Brooks** is an assistant professor in environmental studies at the University of Colorado, Boulder. **Amanda E. Cravens** is a research social scientist at the US Geological Survey's Fort Collins Science Center in Colorado. **Rebecca L. Nelson** is an associate professor focusing on environmental and natural-resources law at the Melbourne Law School, University of Melbourne, Australia. **Dan R. Reineman** is an assistant professor of environmental science and resource management at California State University Channel Islands in Camarillo. **Nicola Ulibarri** is an assistant professor of urban planning and public policy at the University of California, Irvine.

Where I work
Christin Khan

Photographed by Allison Henry/NOAA/NEFSC.

I am fortunate to spend many of my work days in a small aeroplane with two other scientists and two pilots searching for North Atlantic right whales (*Eubalaena glacialis*). When we spot one, we fly overhead and photograph it from above. That angle lets us identify individual whales, which is important in a species this rare: there are thought to be only 409 animals left, according to the North Atlantic Right Whale Consortium (NARWC), a data-sharing group.

The aircraft, a De Havilland Twin Otter, is owned by my employer, the US National Oceanic and Atmospheric Administration. We need to fly 'low and slow' over the ocean to spot and photograph the whales, so we fly at just 305 metres. The aircraft's design also lets us fly 'slow' at 185 kilometres per hour. The view from the bubble window under the high wing, where I'm peering from in this picture, is phenomenal.

The photos and data we collect become part of a widely used data set that is shared among researchers in the NARWC. And when we find three or more whales together, outside their protected gathering areas, it triggers a warning for ships to slow down to avoid lethal collisions.

I have never had an emergency in the air, although there have been tragedies in the close-knit right-whale research community. I thought about the risks in 2009, before the birth of my first child. The pandemic has grounded us since 16 March, and I'm having similar thoughts about risk as we talk about resuming aerial surveys. I'm not risking only myself. My exposure could affect my family and my community.

I've devoted my whole career to this species and it's hard to watch them careening towards extinction. I'd rather live life to its fullest and protect these whales from human harm than not take the risk.

Christin Khan is a fishery biologist with the National Oceanic and Atmospheric Administration at its Northeast Fisheries Science Center in Woods Hole, Massachusetts. **Interview by Madeline Bodin**.

spotlight

AFP VIA GETTY

A patrol robot is used to monitor people's temperatures and disinfect wards at a hospital in Shenyang, China.

Mechanical medics to the rescue

Researchers are trying to support China's overburdened medical sector with cutting-edge robots, but progress has been slow. **By Sarah O'Meara**

In the first few months of 2020, the outbreak of COVID-19 in China, where 84,000 people have been infected and 4,600 have died, revealed the country's willingness and readiness to deploy robotic technology as part of a medical emergency. Service robots were used in hospitals and publicly shared spaces to clean, take temperatures and deliver food, to minimize contact between people as part of the fight against the coronavirus.

"I was staying in a hotel under quarantine and had my takeaway food delivered by a white, cylindrical robot on wheels with a screen on top and a digitally lockable hatch for food placed inside," says Guang-Zhong Yang, founding dean of the Institute of Medical Robotics at Shanghai Jiao Tong University – China's first academic establishment dedicated to the study of medical robotics, which opened in 2019.

"I ordered the food from the restaurant by phone, it was delivered and the robot brought it to my room and rang my phone, so I could open the door and take the food," he explains.

"In the United Kingdom, I would have classed that kind of robot activity as a novelty, but in China it's gradually becoming less unusual," says Yang, who moved to Shanghai last year after working as director of the Hamlyn Centre for Robotic Surgery at Imperial College London for 12 years.

This gear change in the use of robotics began in 2012, when China's five-year economic plan, published as a statement of intent by the central government, made it clear that service robots would become a key technology. The idea was to make them capable of performing a range of crucial social functions, from firefighting to minimally invasive surgery.

The use of robots in the medical sector, to help in areas such as nursing, physical rehabilitation and surgery, has been a particular priority, says Yao Li, a biomedical and robotics engineer at Stanford Robotics Laboratory in California and founder of Borns

spotlight

Medical Robotics, based in Chengdu, China, and Silicon Valley, California.

"China's need for skilled clinical staff in areas such as health care has contributed to the government's focus on the robotics industry to help care for citizens in the future," he says.

The country currently has a shortage of health-care workers. In 2017, China had 2 doctors per 1,000 members of its population, whereas the average in countries that are part of the Organisation for Economic Co-operation and Development was 3.5. Evidence shows that increased living standards have sparked a demand for health-care services, and that the country's rapidly ageing population will place more demands on the system in the future (Q. Wu *et al*. *Br. Med. J.* **354**, i4860; 2016).

According to Jian-Kun Hu, director of surgery at the West China Hospital of Sichuan University in Chengdu, one of China's most prestigious medical centres, his hospital began planning the introduction of robotic technology in 2012. The aims were to give patients the benefits of minimally invasive surgery and to reduce some of the heavy workload on staff. For example, during surgery for gastric or colorectal cancer, the robotic system helps surgeons to see small lymphatic vessels, veins and nerves that need to be protected, Hu says. The outbreak of COVID-19 has prompted hospitals to speed up the clinical application of robotic technology, he adds.

"The use of robots in the medical sector, to help in areas such as nursing, has been a particular priority."

In 2015, the hospital purchased a US-made general-surgical system for minimally invasive surgery called da Vinci – a four-armed, chandelier-like apparatus operated by a surgeon through a computer console. That year, 12 other similar systems were installed in the country (see page S51). In 2018, the hospital installed a ROSA robotic surgical assistant for use in neurosurgery and last September acquired a logistics robot, which has been disinfecting isolation wards during the COVID-19 outbreak. This year, the hospital intends to expand the use of logistics robots, to reduce the burden on and danger to staff, he says.

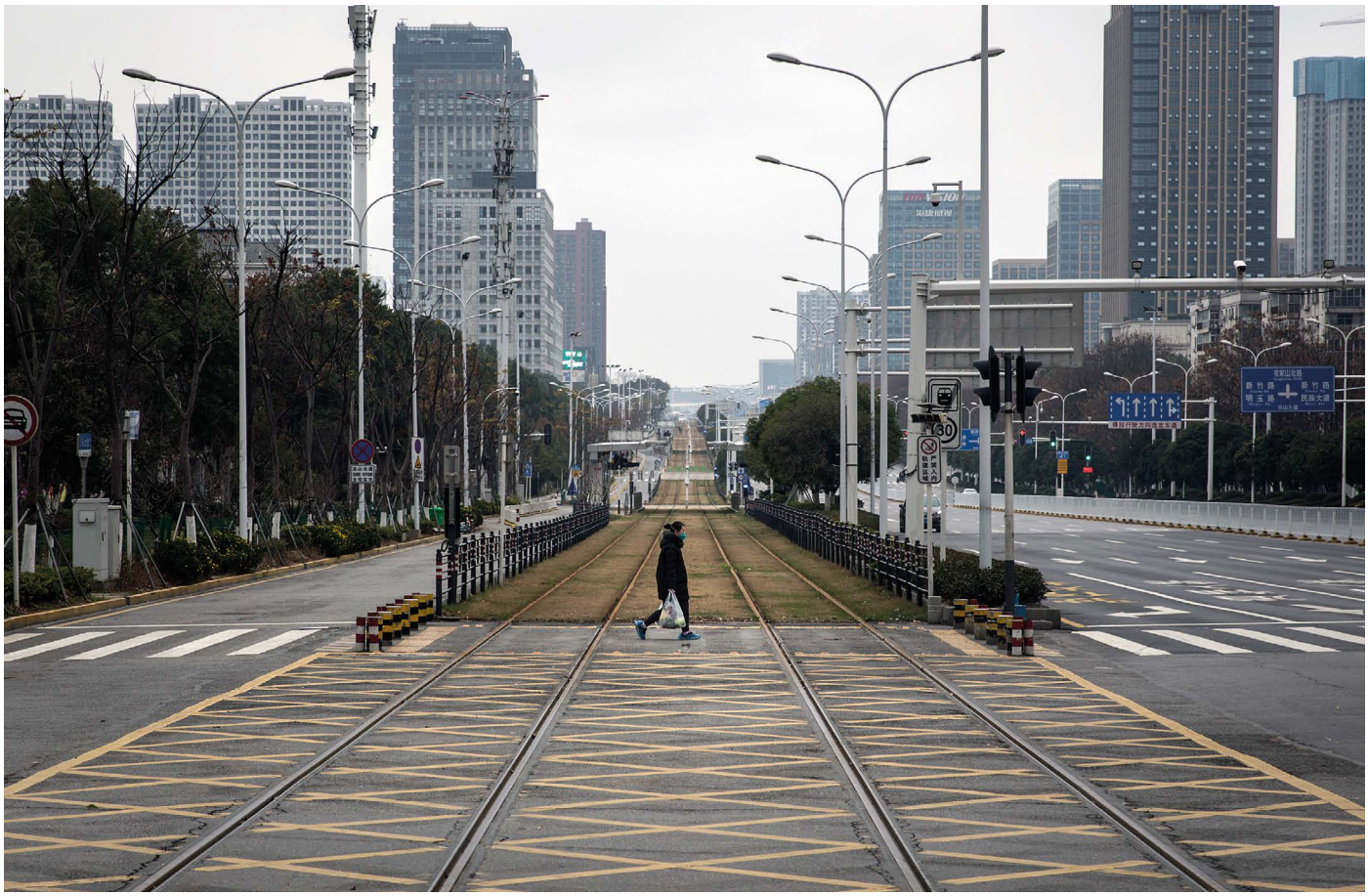
Home-grown robots

Although hospitals are keen to use more robots, the market for such technology in China is relatively young, says Miao Li, the co-founder of Cobot, a four-year-old company in Wuhan that makes easy-to-use operating systems for multipurpose service and industrial robots.

At present, only service robots that do basic jobs such as delivering drugs and food to people are affordable to businesses and hospitals, he says. "You can now buy these simple service robots for around US$10,000 because these robots are also used in hotels, restaurants and other similar scenarios."

A disinfection robot for use in a hospital, however, will usually cost $30,000–80,000, says Li. The da Vinci technology cost the West China hospital $3.5 million, according to Hu.

What the Chinese robotics market needs,



GETTY

The streets of Wuhan in China are deserted during lockdown because of COVID-19, which has spurred efforts to use robots in hospitals.

says Yao Li, is an increased number of home-grown Chinese robotics companies that will stimulate competition and demand, and ultimately lower costs.

China does not make its own equivalent of the da Vinci system, but it is starting to catch up. In 2016, Beijing-based company Tinavi Medical Technologies received fast-tracked approval from the central government to sell the TiRobot, the first robot-assisted surgical product made in China. It has a single arm that can conduct spinal surgery.

Li hopes that his company, which was established in 2016, will soon launch its robotic surgical system. He plans to make an integrated robotic product that runs on a custom-built software platform, comprising robotic hardware — including dexterous multi-jointed arms — and electronics that work with conventional surgical instruments, such as endoscopes, giving them more movability, precision and stability. Unlike other robotic systems, such as the da Vinci, the Borns platform will also act as a data centre and record information about operations that have been performed, says Li.

"We will use the data we collect from surgical procedures to improve the performance of the platform," he adds.

However, he says that the production process has been more challenging than he expected. Li is now working towards gaining the registration his company needs to start production, and predicts it could take a few more steps to bring his robot to the market.

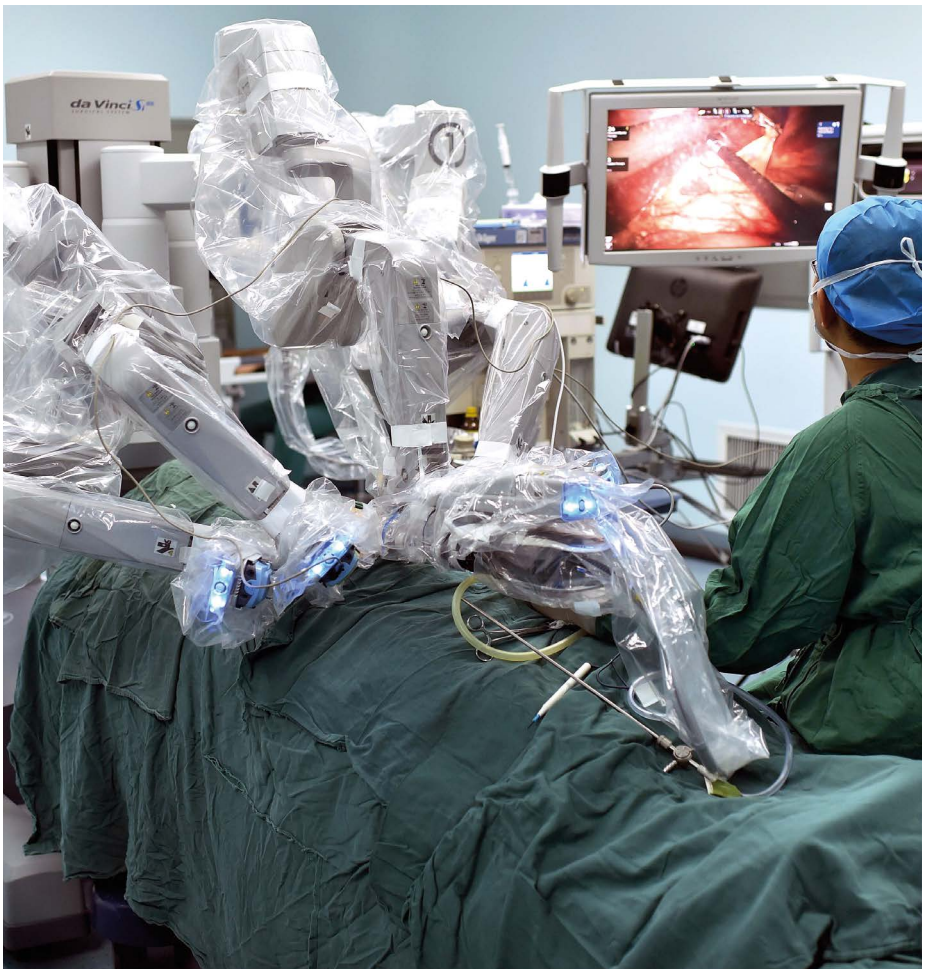
He says developing a product that fully meets the needs of surgeons, hospital staff and patients, while being affordable and also satisfying government regulations and purchasing requirements, is a difficult and lengthy process.

Despite going through a rigorous, three-phase process that includes certification of the technology's functionality, validation of its clinical use and obtaining a licence for its manufacture, Li says he hopes to have the whole process wrapped up within five years.

Li founded his company, which is funded entirely by private and venture-capital funds, with computer-systems engineer William Levine. They met when Li was a doctoral student at the University of Maryland in College Park.

"You have to think about surgeons, hospitals and patients as you're developing your product. If you can keep all those things balanced then you can keep your technology alive," says Li.

Sarah O'Meara is a freelance journalist based in London. Additional reporting by Kevin Schoenmakers.



SHUTTERSTOCK

A da Vinci surgical robot system performs heart surgery in 2017 at a hospital in Hefei, China.

MEDICAL ROBOTICS ON THE RISE

China's push to get more robots into its hospitals is starting to bear fruit. **By Sarah O'Meara**

In 2006, China highlighted the importance of robotics in its 15-year plan for science and technology. In 2011, the central government fleshed out these ambitions in its 12th five-year plan, specifying that robots should be used to support society in a wide range of roles, from helping emergency services during natural disasters and firefighting, to performing complex surgery and aiding in medical rehabilitation.

Guang-Zhong Yang, head of the Institute of Medical Robotics at Shanghai Jiao Tong University, says that China's robotics research output has been growing steadily for two decades, driven by three major factors: "The clinical utilization of robotics; increased funding levels driven by national planning needs; and advances in engineering in areas such as precision mechatronics, medical imaging, artificial intelligence and new materials for making robots."

Yang points out that funding levels for medical robotics from the National Natural Science Foundation of China and the Ministry of Science and Technology began to increase more sharply in 2011 compared to the previous decade.

says Yao Li, is an increased number of home-grown Chinese robotics companies that will stimulate competition and demand, and ultimately lower costs.

China does not make its own equivalent of the da Vinci system, but it is starting to catch up. In 2016, Beijing-based company Tinavi Medical Technologies received fast-tracked approval from the central government to sell the TiRobot, the first robot-assisted surgical product made in China. It has a single arm that can conduct spinal surgery.

Li hopes that his company, which was established in 2016, will soon launch its robotic surgical system. He plans to make an integrated robotic product that runs on a custom-built software platform, comprising robotic hardware — including dexterous multi-jointed arms — and electronics that work with conventional surgical instruments, such as endoscopes, giving them more movability, precision and stability. Unlike other robotic systems, such as the da Vinci, the Borns platform will also act as a data centre and record information about operations that have been performed, says Li.

"We will use the data we collect from surgical procedures to improve the performance of the platform," he adds.

However, he says that the production process has been more challenging than he expected. Li is now working towards gaining the registration his company needs to start production, and predicts it could take a few more steps to bring his robot to the market.

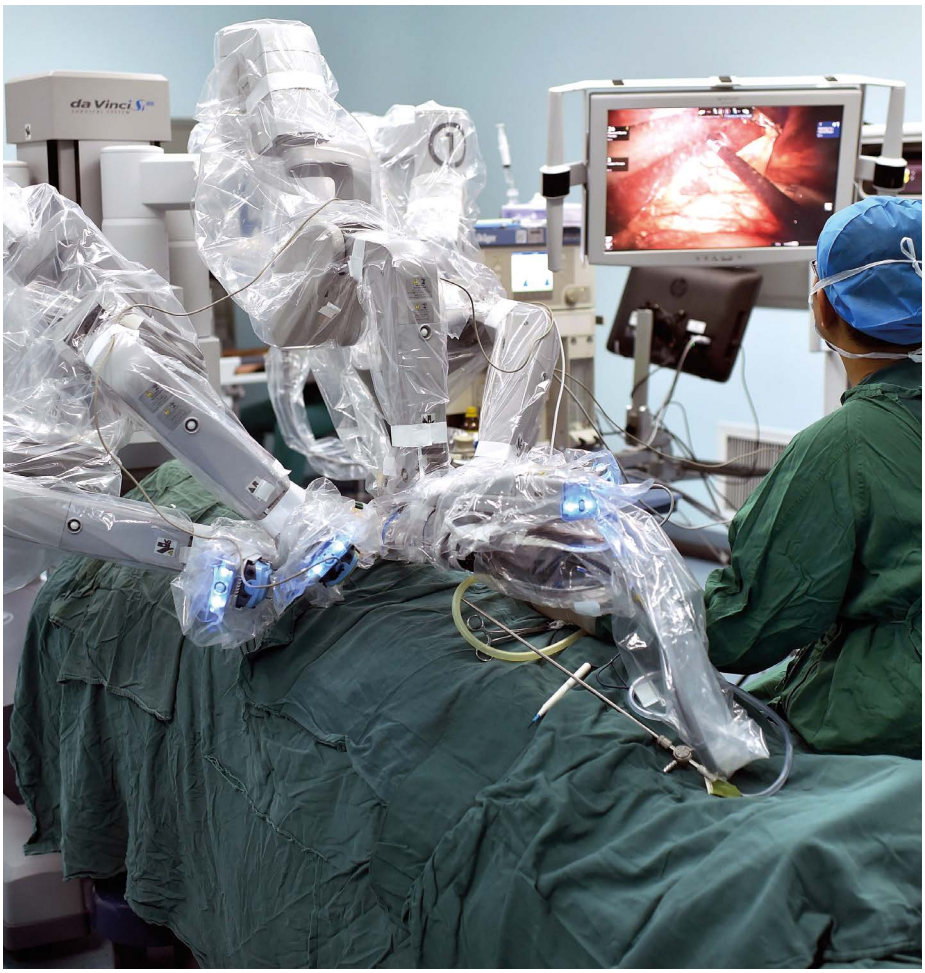
He says developing a product that fully meets the needs of surgeons, hospital staff and patients, while being affordable and also satisfying government regulations and purchasing requirements, is a difficult and lengthy process.

Despite going through a rigorous, three-phase process that includes certification of the technology's functionality, validation of its clinical use and obtaining a licence for its manufacture, Li says he hopes to have the whole process wrapped up within five years.

Li founded his company, which is funded entirely by private and venture-capital funds, with computer-systems engineer William Levine. They met when Li was a doctoral student at the University of Maryland in College Park.

"You have to think about surgeons, hospitals and patients as you're developing your product. If you can keep all those things balanced then you can keep your technology alive," says Li.

Sarah O'Meara is a freelance journalist based in London. Additional reporting by Kevin Schoenmakers.



SHUTTERSTOCK

A da Vinci surgical robot system performs heart surgery in 2017 at a hospital in Hefei, China.

MEDICAL ROBOTICS ON THE RISE

China's push to get more robots into its hospitals is starting to bear fruit. **By Sarah O'Meara**

In 2006, China highlighted the importance of robotics in its 15-year plan for science and technology. In 2011, the central government fleshed out these ambitions in its 12th five-year plan, specifying that robots should be used to support society in a wide range of roles, from helping emergency services during natural disasters and firefighting, to performing complex surgery and aiding in medical rehabilitation.

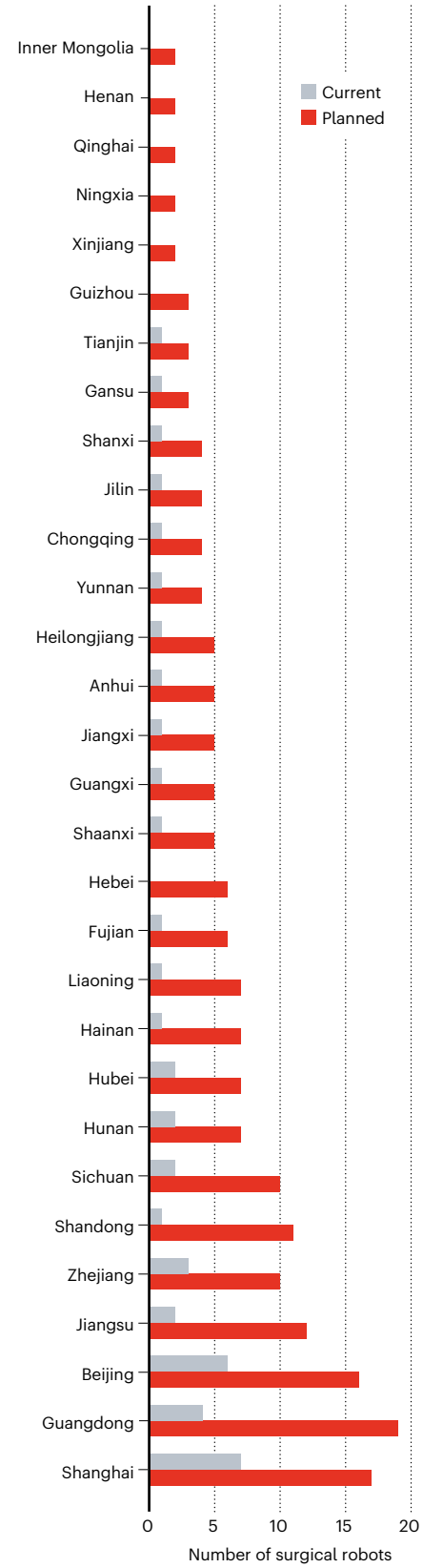
Guang-Zhong Yang, head of the Institute of Medical Robotics at Shanghai Jiao Tong University, says that China's robotics research output has been growing steadily for two decades, driven by three major factors: "The clinical utilization of robotics; increased funding levels driven by national planning needs; and advances in engineering in areas such as precision mechatronics, medical imaging, artificial intelligence and new materials for making robots."

Yang points out that funding levels for medical robotics from the National Natural Science Foundation of China and the Ministry of Science and Technology began to increase more sharply in 2011 compared to the previous decade.

spotlight

SURGICAL ROBOTS ACROSS CHINA

Hospitals in mainland China's administrative areas have big plans for installing more surgical robotics systems.



The accompanying rises in research output are closely related to the introduction of specialized robotics equipment in medical-research facilities, says Yao Li, a research scientist at Stanford Robotics Laboratory in California and founder of the company Borns Medical Robotics, based in both Chengdu, China, and Silicon Valley, California.

> **"Robots should be used to support society in a wide range of roles from firefighting to surgery."**

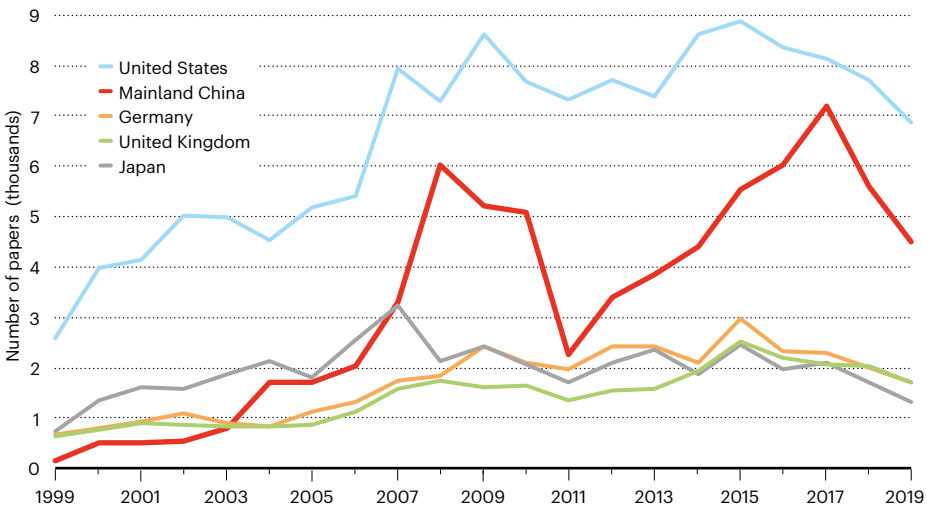
Between 1999 and 2019, the number of papers published by at least one Chinese author in the combined fields of biomedical engineering and robotics increased from 142 to 4,507, and spiked twice during that period (see 'Published papers'), according to data from the Web of Science. One peak was in 2008, two years after a robotic system for minimally invasive operations called da Vinci was first deployed to hospitals in China. The second was in 2017, a year after the first Chinese-designed robot for minimally invasive spinal surgery was approved for sale.

In 2019, the number of da Vinci systems installed in Chinese hospitals that year leapt to 59, up from only 8 installations in 2018 (see 'Spike in hospital robotics'). This surge followed a 2018 government push to encourage research on robotics technology and its clinical application, according to Jian-Kun Hu, director of the department of gastrointestinal surgery at West China Hospital in Chengdu. The central government's plan included an intention to purchase 154 new surgical robot systems by the end of 2020, and a breakdown of how the systems would be allocated nationwide (see 'Surgical robots across China').

Sarah O'Meara is freelance journalist based in London.

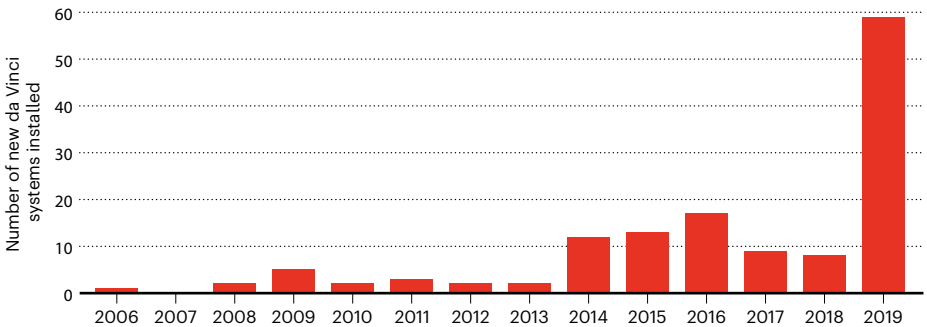
PUBLISHED PAPERS

China has stepped up research efforts into medical robotics systems in the past two decades, as shown by the number of papers it published in biomedical robotics. Papers in this field from the United States have also risen, whereas those from other countries have remained steady.



SPIKE IN HOSPITAL ROBOTICS

Last year saw a large jump in the number of da Vinci robotics systems installed in Chinese hospitals.

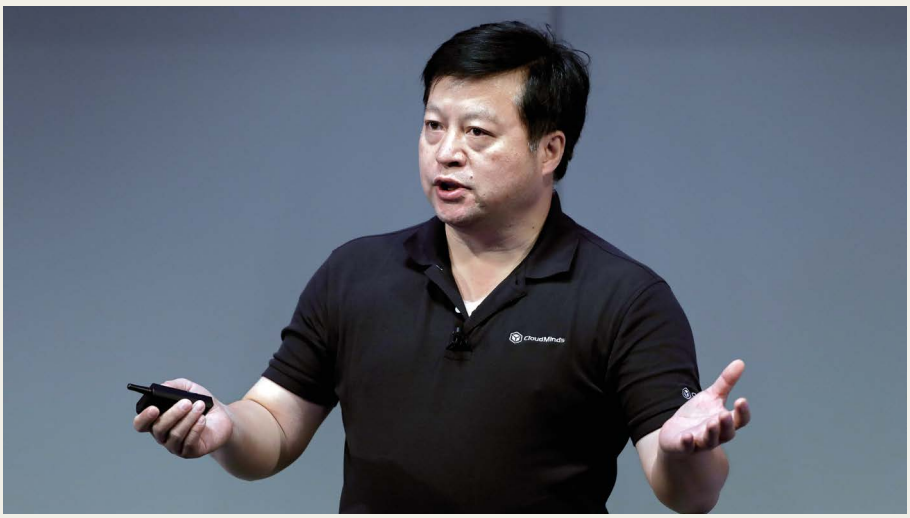


SOURCES: SURGICAL ROBOTS ACROSS CHINA: NATL HEALTH COMMISSION, PRC; PAPERS: WEB OF SCIENCE; SPIKE IN HOSPITAL ROBOTICS: JIAN-KUN HU/INTUITIVE SURGICAL–FOSUN MEDICAL TECHNOLOGY (SHANGHAI)

Q&A

Bill Huang
Robot engineer

Bill Huang is chief executive of CloudMinds, a Chinese–US company that provides a cloud-based robot operating system. He explains how the COVID-19 pandemic has accelerated his firm's application of robotics technology in real-life situations.



KIYOSHI OTA/BLOOMBERG VIA GETTY

Bill Huang is the founder and chief executive of CloudMinds, a robotics company in Beijing.

What does your company do?

I founded CloudMinds in 2015 with the long-term goal of making humanoid robot nannies that could be used to help parents who are struggling with the demands of running their home while working full-time.

These kinds of robot are extremely complicated to make because they require sophisticated artificial intelligence to perform a wide range of duties — such as feeding and dressing children.

In the short term, we began by focusing on making robots for three industries: hospitality, retail and health care. We currently produce four types of robot. Humanoid robots for reception work; security patrol robots to guard empty private land or offices; cleaning robots; and intelligent vending machines that can see and talk to you — a popular feature in China and southeast Asia.

How did the outbreak affect your work?

On 24 January, the day after the Chinese government imposed a lockdown on the city of Wuhan, we shifted our company's focus to help to fight the spread of the virus. Since the outbreak, all Chinese robot companies that I know of have done this. Within weeks, we were providing proposals to hospitals for robots that could disinfect wards, monitor temperatures and dispense drugs.

On 28 February, we began setting up China's first robot-run ward, which was designed to prevent staff at Wuhan Wuchang Hospital from contracting COVID-19. It was operational by 6 March, but just four days later it was suspended, because cases of the virus had dropped so sharply across the city.

It was a big experiment for us. As far as I know, there has never been a fully automated, robot-run medical ward in China. We had just weeks to organize its design, reach out to other companies to provide extra equipment and find ways to transport and ship the hardware and necessary engineers to the hospital during lockdown in Wuhan as the rest of China was closing.

How did the robot ward work?

The ward was set up in Wuhan Hongshan Sports Centre. We designed it to be suitable for around 200 people who were demonstrating early symptoms of the virus and needed medical care but were not seriously ill.

In total, we had 12 robots performing different functions, ranging from the delivery of food, drinks and drugs, to monitoring vital signs and disinfecting the area. Outside the room, clinical staff controlled and monitored the robots' movements.

Before we opened the ward, engineers from CloudMinds mapped the area and uploaded its geographical information to a cloud-based server. The robots used these information points to navigate the ward. Our engineers also installed a large screen outside the ward that displayed the health information of each individual, and was used by doctors and nurses to assign the robots to their next task.

Patients were also given bracelets fitted with sensors to measure their heart rates and temperatures.

What were the risks involved?

Our two major concerns were connectivity and technical malfunction. We were less concerned about people's reaction to them because in my experience, people tend not to find robots scary. We've actually found that people like the robots because they're fun to look at and relieve boredom.

All our robots are connected to a central cloud brain by individual, secure virtual networks that run on a mobile operating system. We rent robots to customers, along with operating software, and connect the units to the cloud, where engineers provide technical support. However, if the robots lose their mobile signal then they stop working. Before I started this company, I was a telecommunications engineer and head of research for China Mobile, the world's largest mobile operator. We made an agreement that if the 4G signal dropped, it would be boosted by China Mobile.

Happily, there were no hardware failures and the hospital's director said that if the same kind of event happened in the future, he wouldn't hesitate to roll it out again. That was nice to hear.

Where are the robots now?

They're resting back at our company's headquarters in Beijing. On the basis of the feedback from medical staff and patients, my team is doing further research and development to improve the robots in case they're needed again.

Interview by Sarah O'Meara.

This interview has been edited for length and clarity.